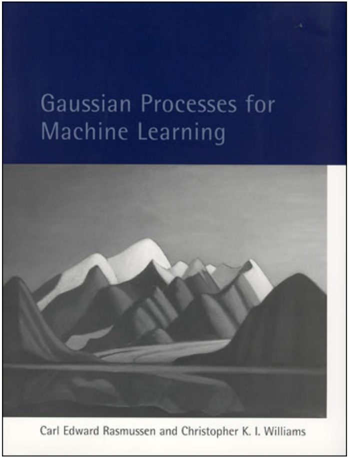
Session 1: Gaussian Processes

Neil D. Lawrence and Raquel Urtasun

CVPR
16th June 2012

Book


Gaussian Processes for
Machine Learning
Carl Edward Rasmussen and Christopher K. I. Williams

?

Outline

Outline

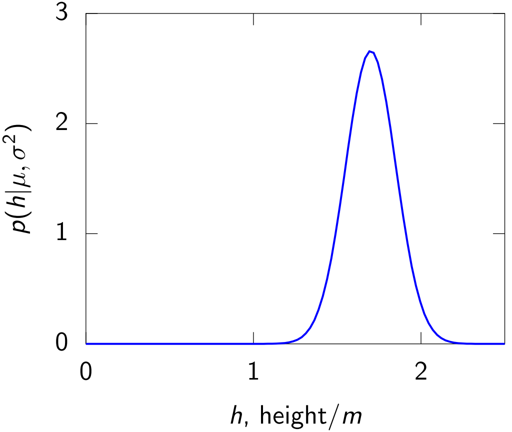
The Gaussian Density

- Perhaps the most common probability density.

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$
$$= \mathcal{N}\left(y|\mu, \sigma^2\right)$$

- The Gaussian density.

Gaussian Density



The Gaussian PDF with $\mu = 1.7$ and variance $\sigma^2 = 0.0225$. Mean shown as red line. It could represent the heights of a population of students.

Gaussian Density

$$\mathcal{N}\left(y|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Two Important Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Two Important Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Two Important Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Two Important Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Two Important Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

   (*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Two Important Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Two Simultaneous Equations

A system of two differential equations with two unknowns.

$$
\begin{aligned}
y_1 &= mx_1 + c \\
y_2 &= mx_2 + c
\end{aligned}
$$

Two Simultaneous Equations

A system of two differential equations with two unknowns.

$$y_1 - y_2 = m(x_1 - x_2)$$

Two Simultaneous Equations

A system of two differential equations with two unknowns.

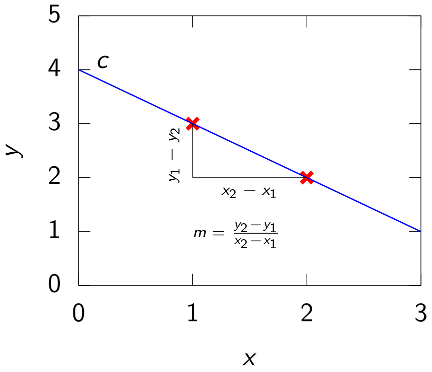
$$\frac{y_1 - y_2}{x_1 - x_2} = m$$

Two Simultaneous Equations

A system of two differential equations with two unknowns.

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

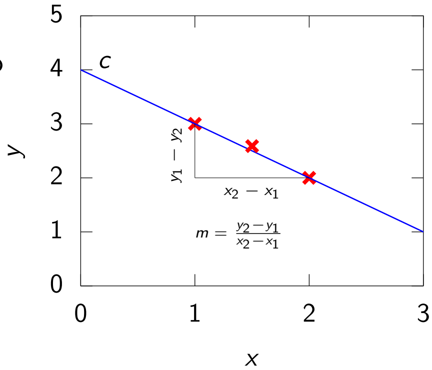
$$c = y_1 - mx_1$$



Two Simultaneous Equations

How do we deal with three simultaneous equations with only two unknowns?

$$
\begin{aligned}
y_1 &= mx_1 + c \\
y_2 &= mx_2 + c \\
y_3 &= mx_3 + c
\end{aligned}
$$



Overdetermined System

- With two unknowns and two observations:

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

- Additional observation leads to *overdetermined* system.

$$y_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$y_1 = mx_1 + c + \epsilon_1$$
$$y_2 = mx_2 + c + \epsilon_2$$
$$y_3 = mx_3 + c + \epsilon_3$$

Overdetermined System

- With two unknowns and two observations:

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

- Additional observation leads to *overdetermined* system.

$$y_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$y_1 = mx_1 + c + \epsilon_1$$
$$y_2 = mx_2 + c + \epsilon_2$$
$$y_3 = mx_3 + c + \epsilon_3$$

Overdetermined System

- With two unknowns and two observations:

$$
\begin{aligned}
y_1 =& mx_1 + c \\
y_2 =& mx_2 + c
\end{aligned}
$$

- Additional observation leads to *overdetermined* system.

$$y_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$
\begin{aligned}
y_1 &= mx_1 + c + \epsilon_1 \\
y_2 &= mx_2 + c + \epsilon_2 \\
y_3 &= mx_3 + c + \epsilon_3
\end{aligned}
$$

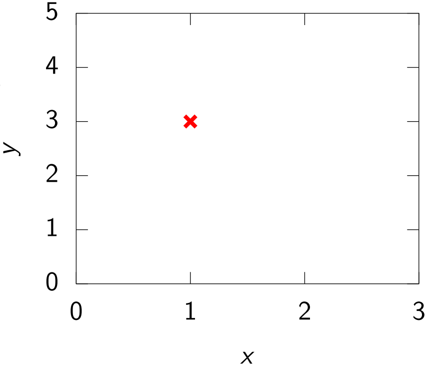
Noise Models

- We aren't modeling entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to *least squares*.

Underdetermined System

What about two unknowns and *one* observation?

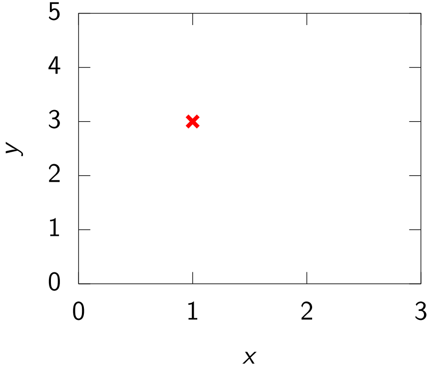
$$y_1 = mx_1 + c$$



Underdetermined System

Can compute m given c.

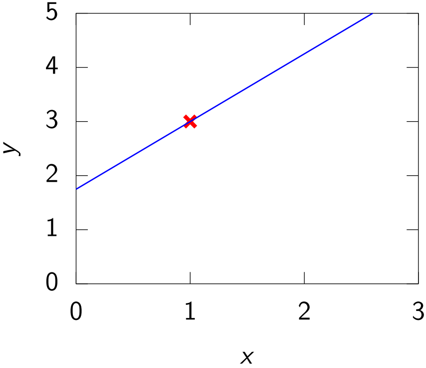
$$m = \frac{y_1 - c}{x}$$



Underdetermined System

Can compute m given c.

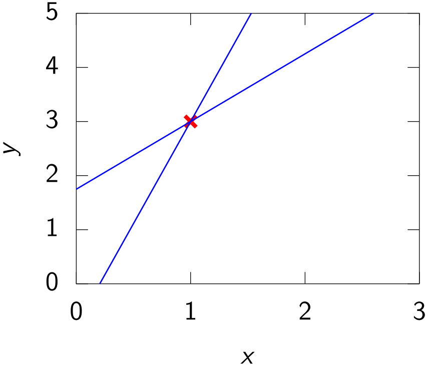
$$c = 1.75 \implies m = 1.25$$



Underdetermined System

Can compute m given c.

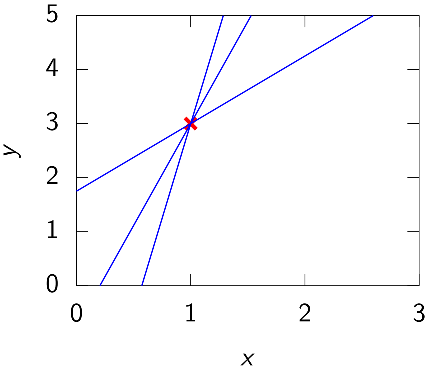
$$c = -0.777 \Longrightarrow m = 3.78$$



Underdetermined System

Can compute m given c.

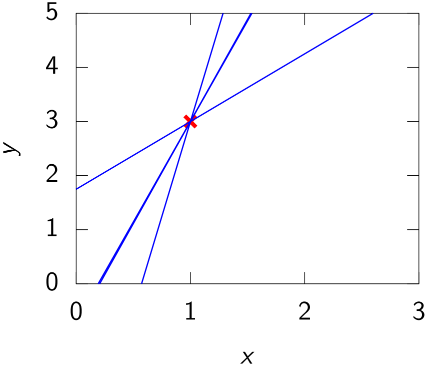
$$c = -4.01 \implies m = 7.01$$



Underdetermined System

Can compute m given c.

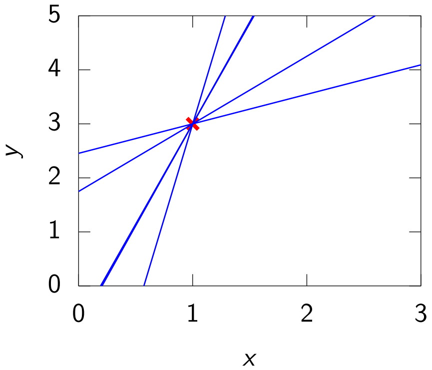
$$c = -0.718 \Longrightarrow m = 3.72$$



Underdetermined System

Can compute m given c.

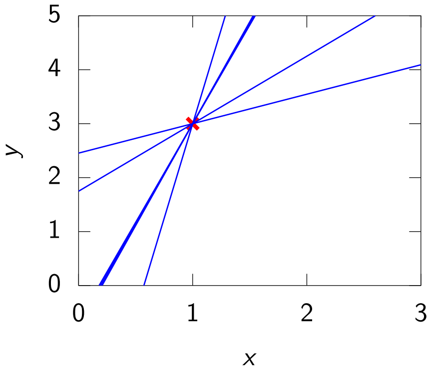
$$c = 2.45 \Longrightarrow m = 0.545$$



Underdetermined System

Can compute m given c.

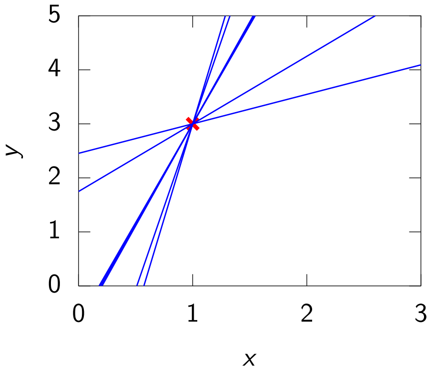
$$c = -0.657 \Longrightarrow m = 3.66$$



Underdetermined System

Can compute m given c.

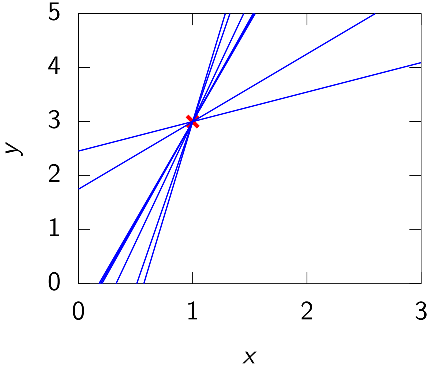
$$c = -3.13 \implies m = 6.13$$



Underdetermined System

Can compute m given c.

$$c = -1.47 \implies m = 4.47$$

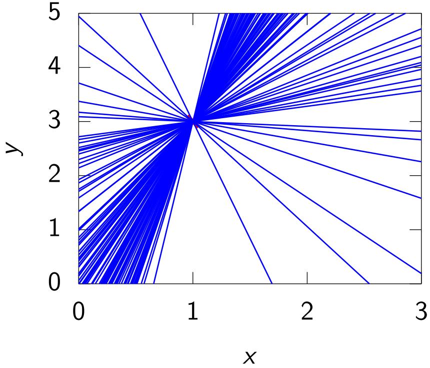


Underdetermined System

Can compute m given c.
Assume

$$c \sim \mathcal{N}(0, 4),$$

we find a distribution of solutions.



Probability for Under- and Overdetermined

- To deal with overdetermined introduced probability distribution for 'variable', ϵ_i.
- For underdetermined system introduced probability distribution for 'parameter', c.
- This is known as a Bayesian treatment.

- For general Bayesian inference need multivariate priors.
- E.g. for multivariate linear regression:

$$y_i = \sum_i w_j x_{i,j} + \epsilon_i$$

(where we've dropped c for convenience), we need a prior over $\mathbf{w}$.
- This motivates a *multivariate* Gaussian density.
- We will use the multivariate Gaussian to put a prior *directly* on the function (a Gaussian process).

- For general Bayesian inference need multivariate priors.
- E.g. for multivariate linear regression:

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

(where we've dropped c for convenience), we need a prior over $\mathbf{w}$.

- This motivates a *multivariate* Gaussian density.
- We will use the multivariate Gaussian to put a prior *directly* on the function (a Gaussian process).

Multivariate Regression Likelihood

- Recall multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_{i,:}\right)^2\right)$$

- Now use a multivariate Gaussian prior:

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

Multivariate Regression Likelihood

- Recall multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_{i,:}\right)^2\right)$$

- Now use a multivariate Gaussian prior:

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

Posterior Density

- Once again we want to know the posterior:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

- And we can compute by completing the square.

$$\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i \mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{x}_{i,:}\mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w} + \text{const.}$$

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_w, \mathbf{C}_w\right)$$

$$\mathbf{C}_w = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1})^{-1} \text{ and } \boldsymbol{\mu}_w = \mathbf{C}_w \sigma^{-2}\mathbf{X}^\top \mathbf{y}$$

Posterior Density

- Once again we want to know the posterior:

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) \propto p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$$

- And we can compute by completing the square.

$$\log p(\mathbf{w}|\mathbf{y},\mathbf{X}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i \mathbf{x}_{i,:}^\top \mathbf{w}$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{x}_{i,:}\mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top\mathbf{w} + \text{const.}$$

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_w, \mathbf{C}_w\right)$$

$$\mathbf{C}_w = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1})^{-1} \text{ and } \boldsymbol{\mu}_w = \mathbf{C}_w \sigma^{-2}\mathbf{X}^\top\mathbf{y}$$

Bayesian vs Maximum Likelihood

- Note the similarity between posterior mean

$$\boldsymbol{\mu}_w = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1})^{-1}\sigma^{-2}\mathbf{X}^\top\mathbf{y}$$

- and Maximum likelihood solution

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Marginal Likelihood is Computed as Normalizer

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X})p(\mathbf{y}|\mathbf{X}) = p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})$$

Marginal Likelihood

- Can compute the marginal likelihood as:

$$p(\mathbf{y}|\mathbf{X}, \alpha, \sigma) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \alpha\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)$$

Two Dimensional Gaussian

- Consider height, h/m and weight, w/kg.
- Could sample height from a distribution:

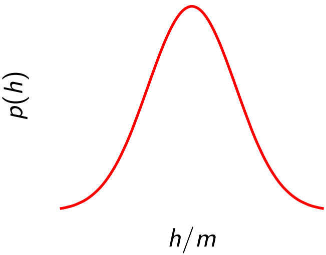
$$p(h) \sim \mathcal{N}(1.7, 0.0225)$$

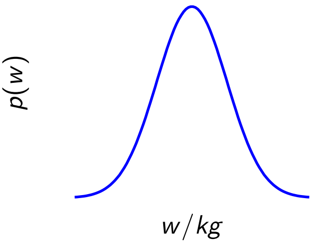
- And similarly weight:

$$p(w) \sim \mathcal{N}(75, 36)$$

Height and Weight Models

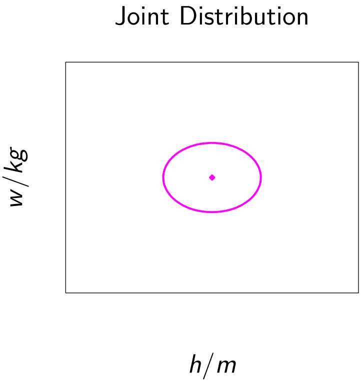
Marginal Distributions

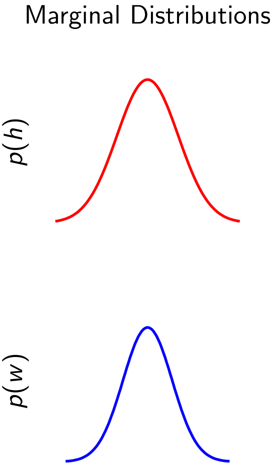




Gaussian distributions for height and weight.

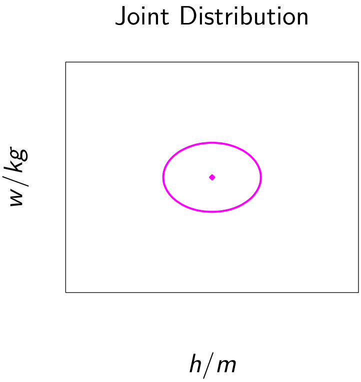
Sampling Two Dimensional Variables

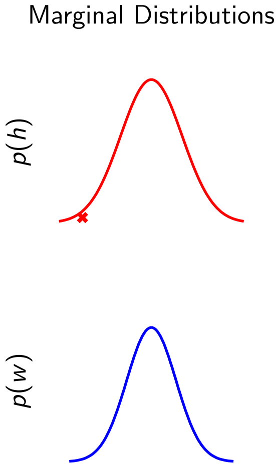




Sample height and weight one after the other and plot against each other.

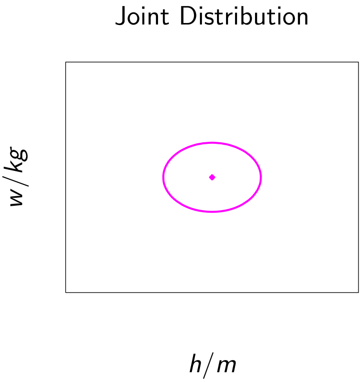
Sampling Two Dimensional Variables

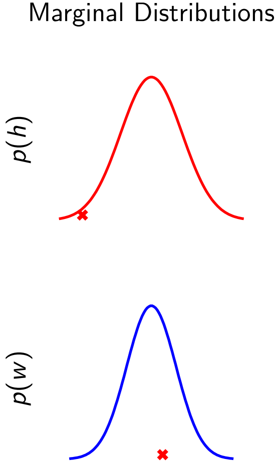




Sample height and weight one after the other and plot against each other.

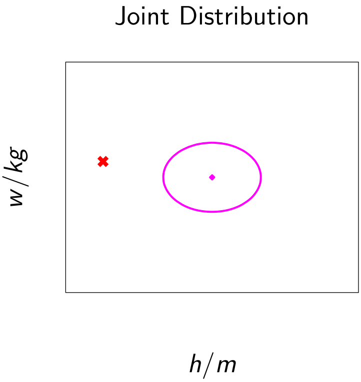
Sampling Two Dimensional Variables

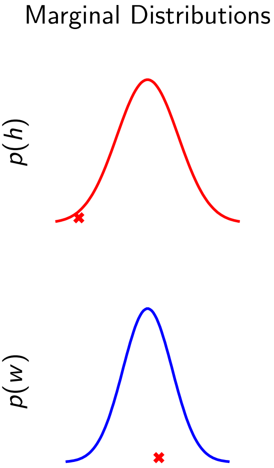




Sample height and weight one after the other and plot against each other.

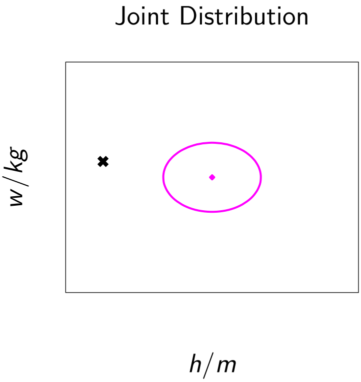
Sampling Two Dimensional Variables

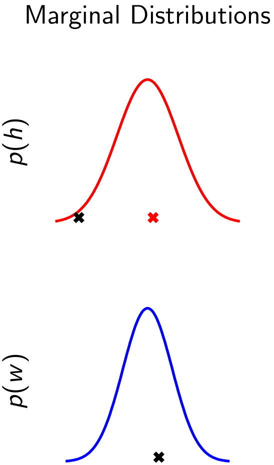




Sample height and weight one after the other and plot against each other.

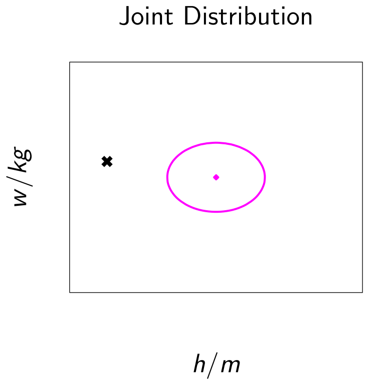
Sampling Two Dimensional Variables

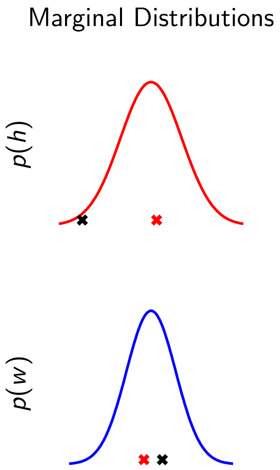




Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables

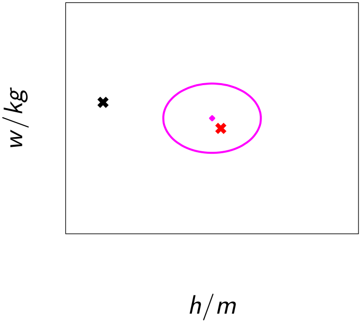




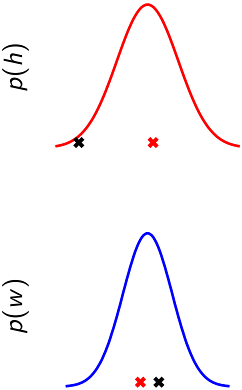
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables

Joint Distribution

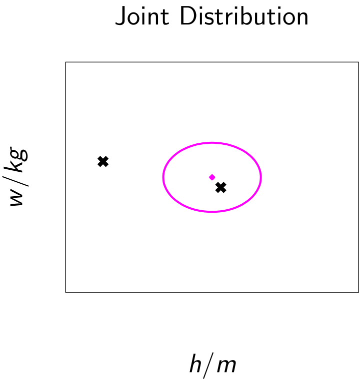


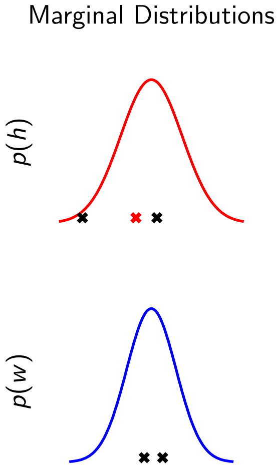
Marginal Distributions



Sample height and weight one after the other and plot against each other.

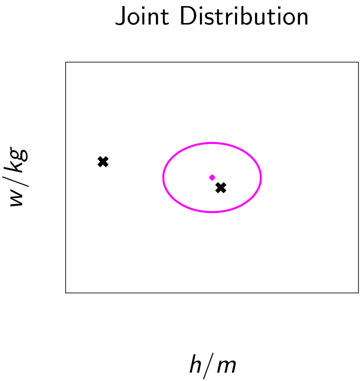
Sampling Two Dimensional Variables

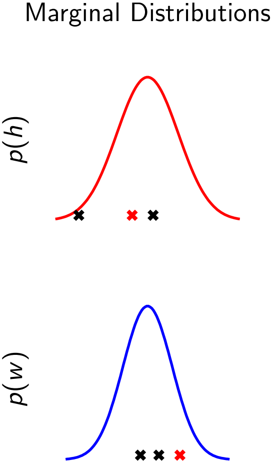




Sample height and weight one after the other and plot against each other.

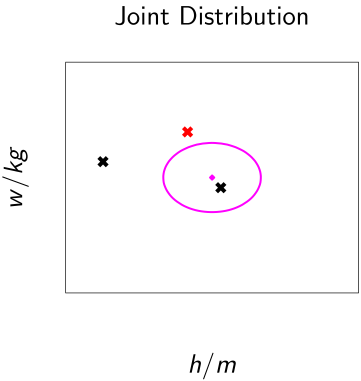
Sampling Two Dimensional Variables

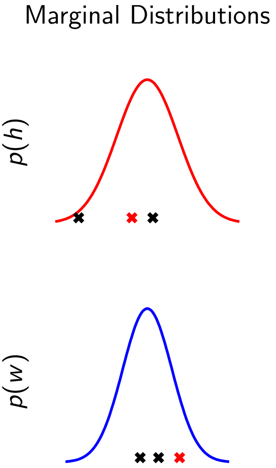




Sample height and weight one after the other and plot against each other.

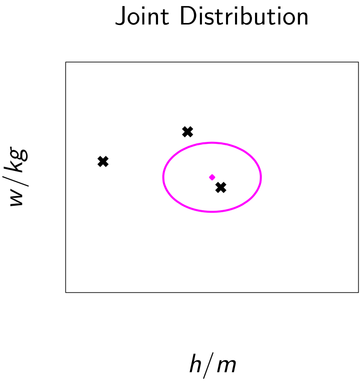
Sampling Two Dimensional Variables

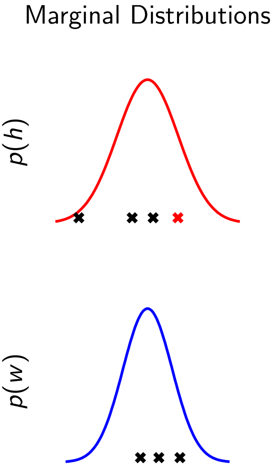




Sample height and weight one after the other and plot against each other.

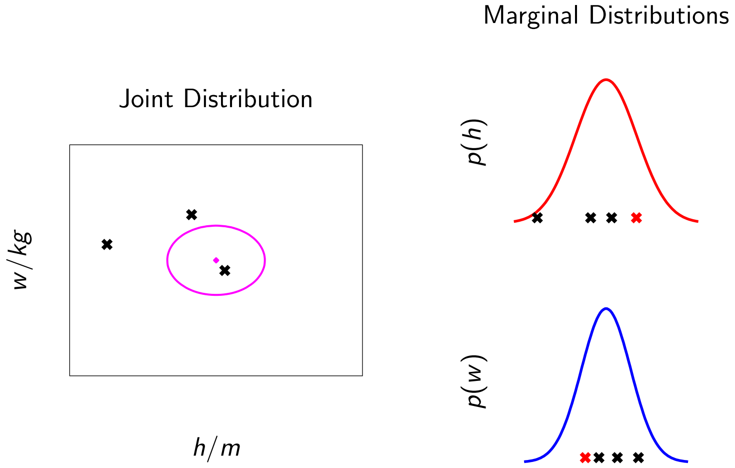
Sampling Two Dimensional Variables





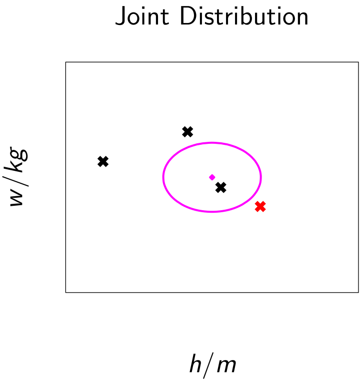
Sample height and weight one after the other and plot against each other.

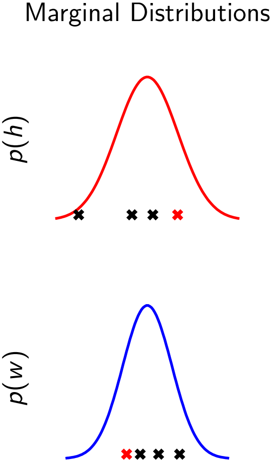
Sampling Two Dimensional Variables



Sample height and weight one after the other and plot against each other.

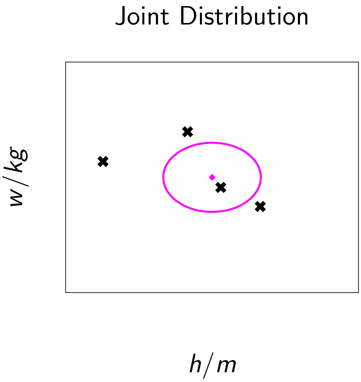
Sampling Two Dimensional Variables

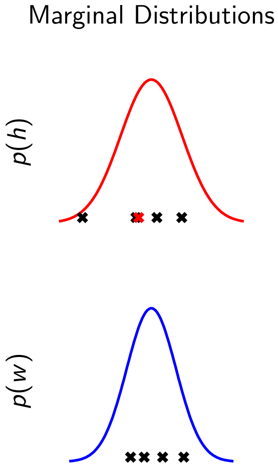




Sample height and weight one after the other and plot against each other.

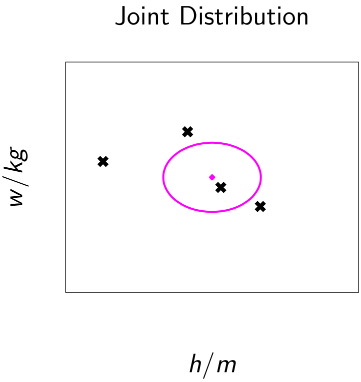
Sampling Two Dimensional Variables

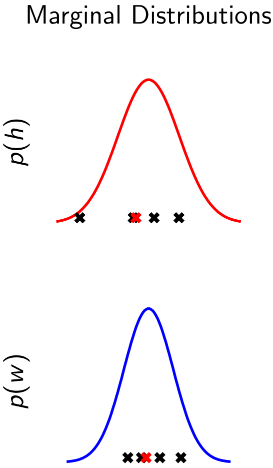




Sample height and weight one after the other and plot against each other.

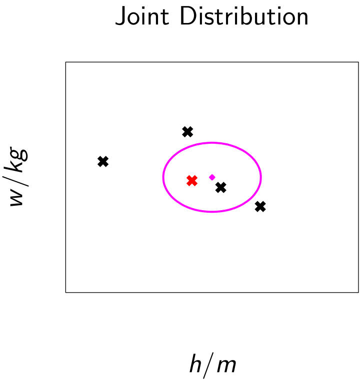
Sampling Two Dimensional Variables

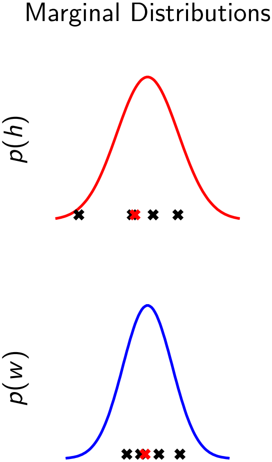




Sample height and weight one after the other and plot against each other.

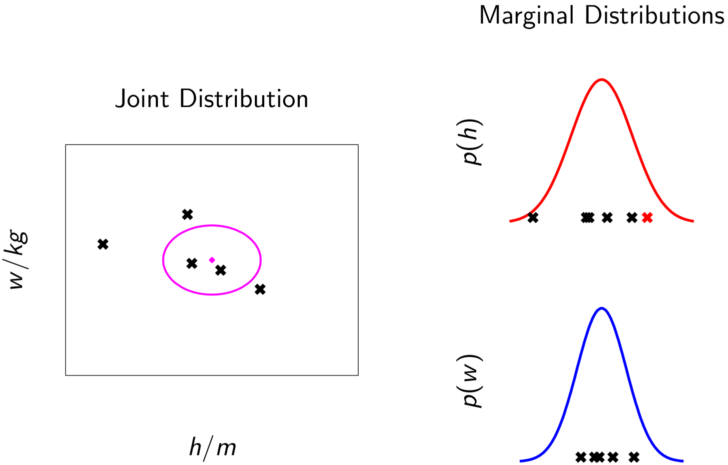
Sampling Two Dimensional Variables





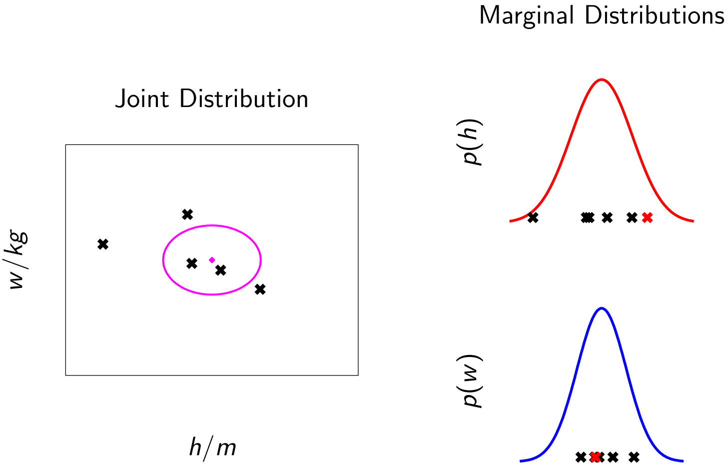
Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables



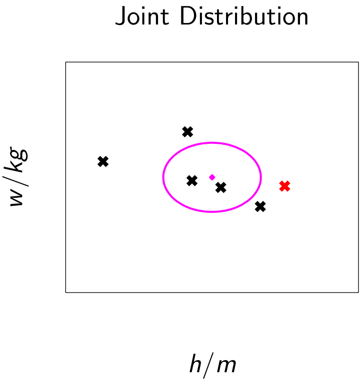
Sample height and weight one after the other and plot against each other.

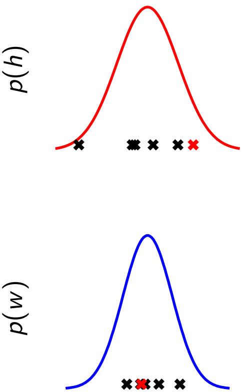
Sampling Two Dimensional Variables



Sample height and weight one after the other and plot against each other.

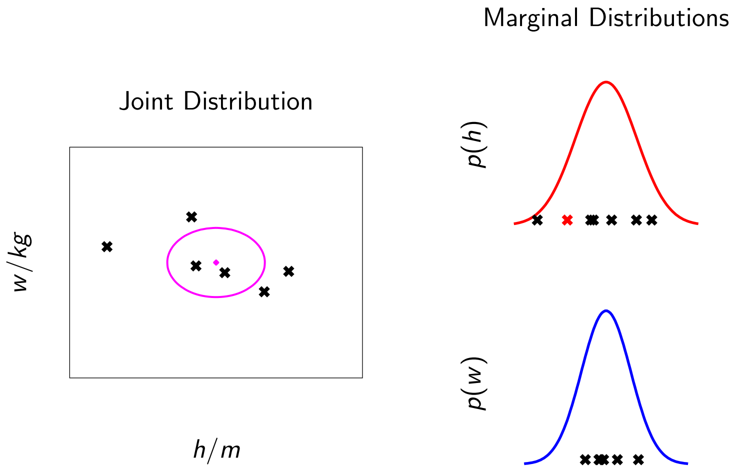
Sampling Two Dimensional Variables





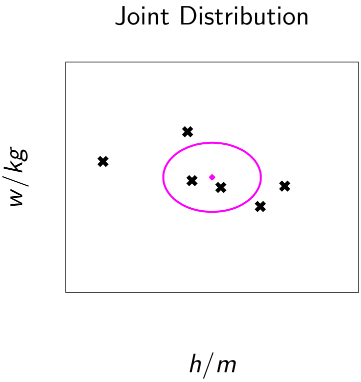
Sample height and weight one after the other and plot against each other.

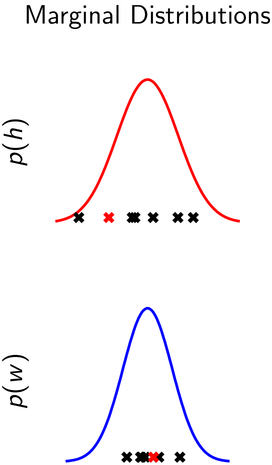
Sampling Two Dimensional Variables



Sample height and weight one after the other and plot against each other.

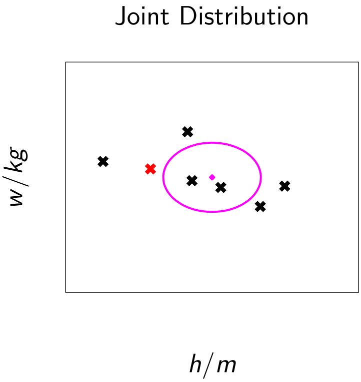
Sampling Two Dimensional Variables

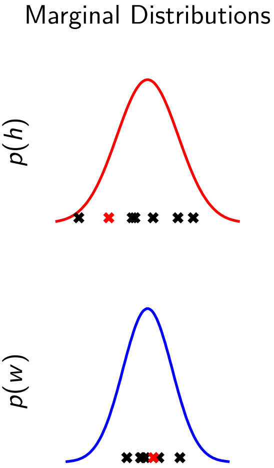




Sample height and weight one after the other and plot against each other.

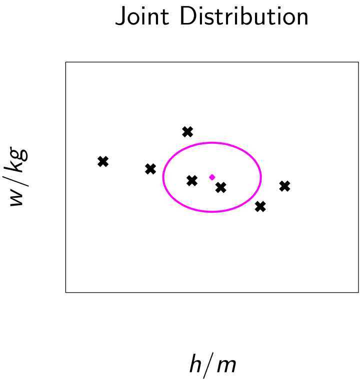
Sampling Two Dimensional Variables

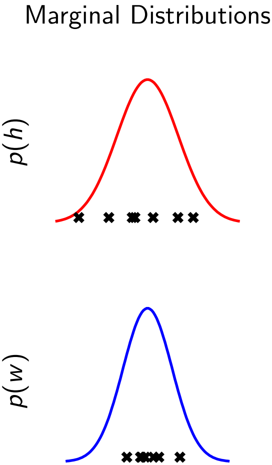




Sample height and weight one after the other and plot against each other.

Sampling Two Dimensional Variables





Sample height and weight one after the other and plot against each other.

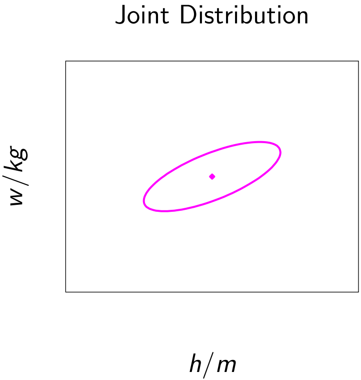
Independence Assumption

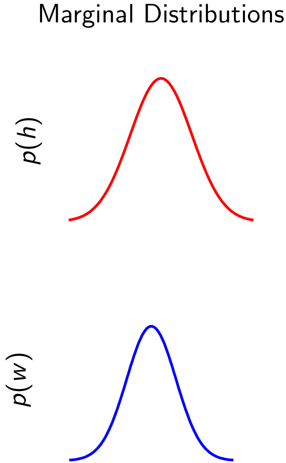
- This assumes height and weight are independent.

$$p(h, w) = p(h)p(w)$$

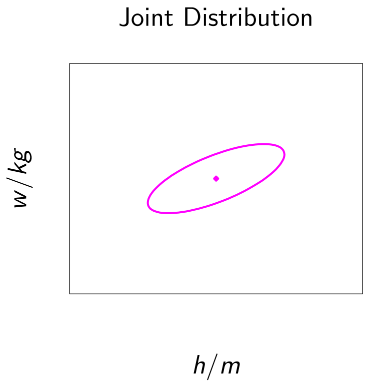
- In reality they are dependent (body mass index) $= \frac{w}{h^2}$.

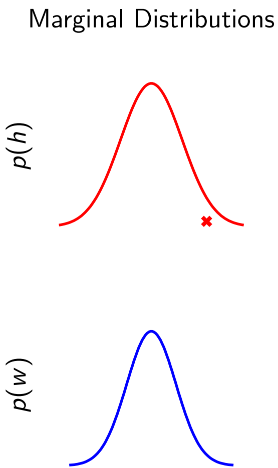
Sampling Two Dimensional Variables



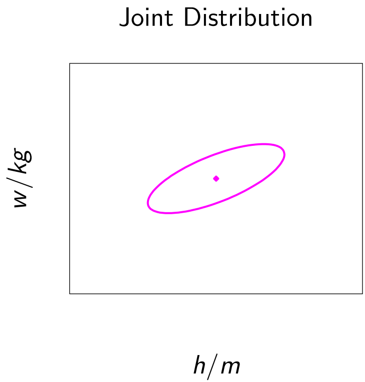


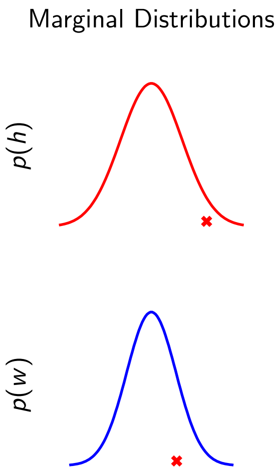
Sampling Two Dimensional Variables



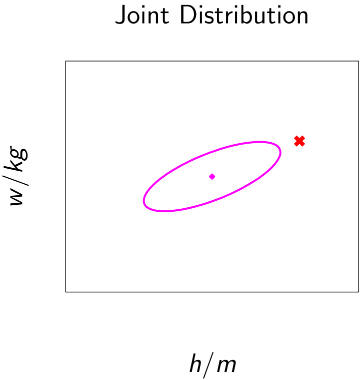


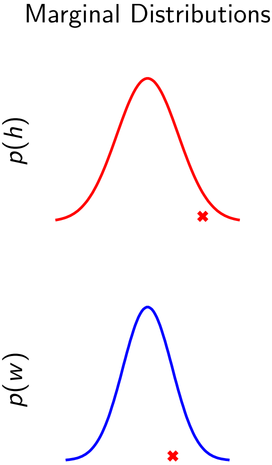
Sampling Two Dimensional Variables



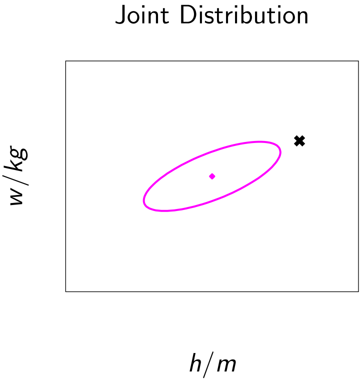


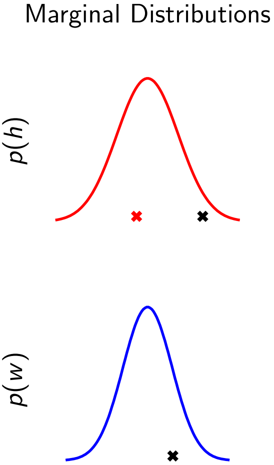
Sampling Two Dimensional Variables





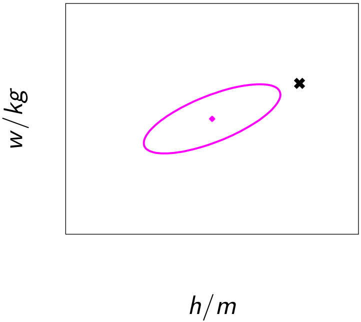
Sampling Two Dimensional Variables



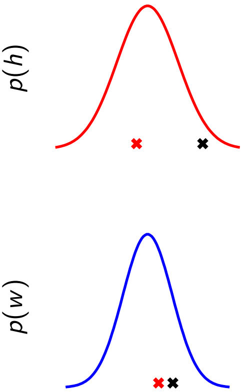


Sampling Two Dimensional Variables

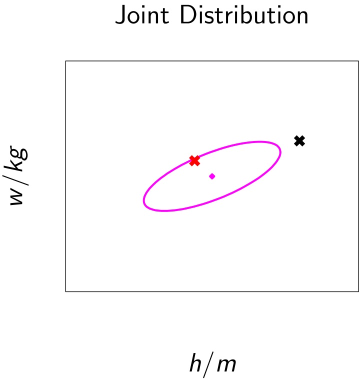
Joint Distribution

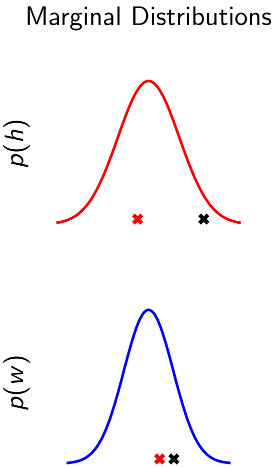


Marginal Distributions

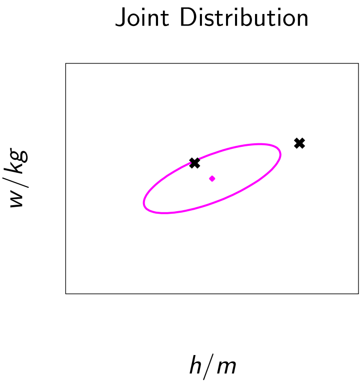


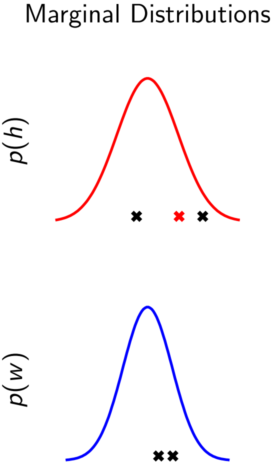
Sampling Two Dimensional Variables



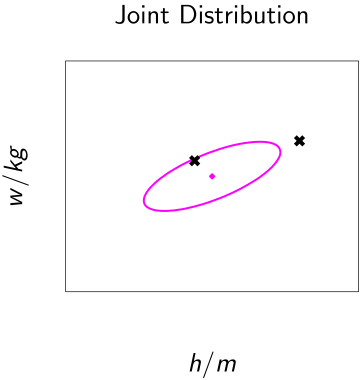


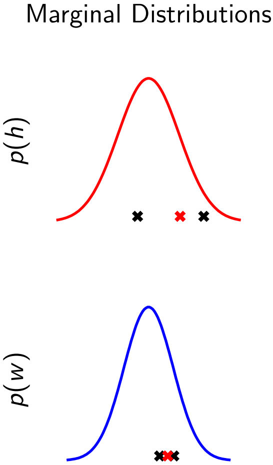
Sampling Two Dimensional Variables



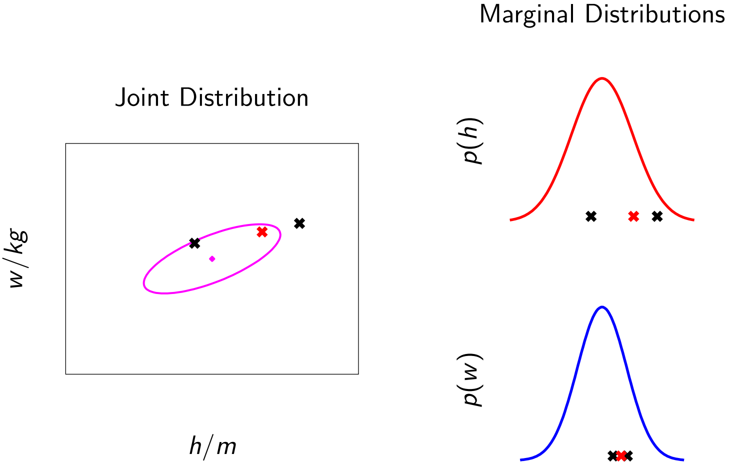


Sampling Two Dimensional Variables

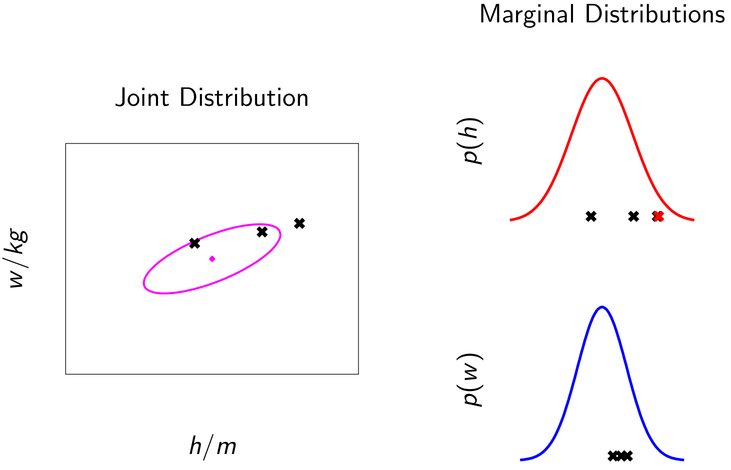




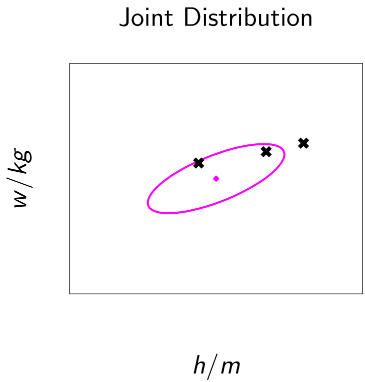
Sampling Two Dimensional Variables

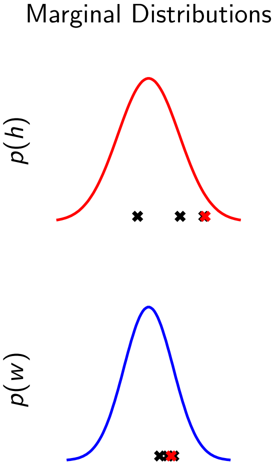


Sampling Two Dimensional Variables

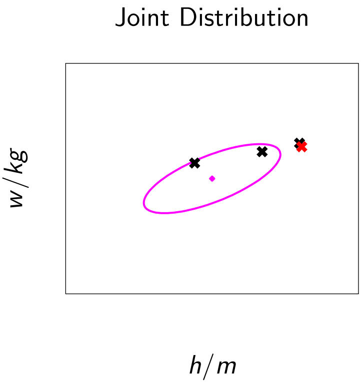


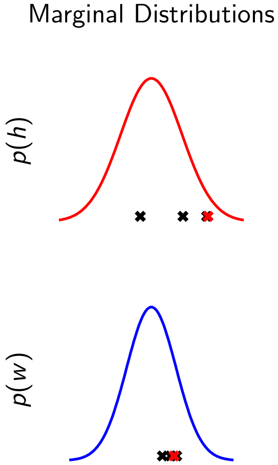
Sampling Two Dimensional Variables



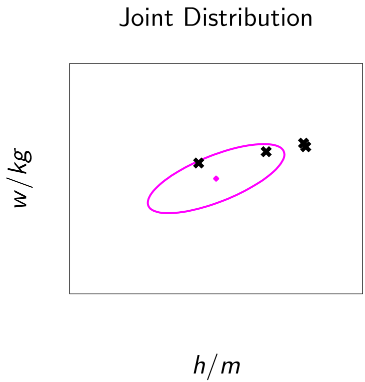


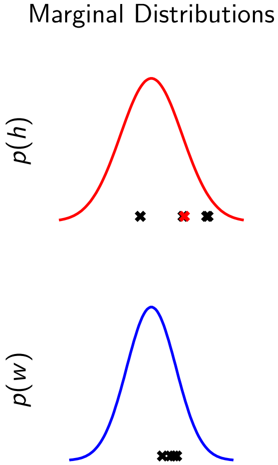
Sampling Two Dimensional Variables



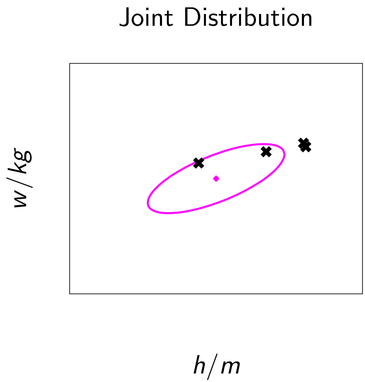


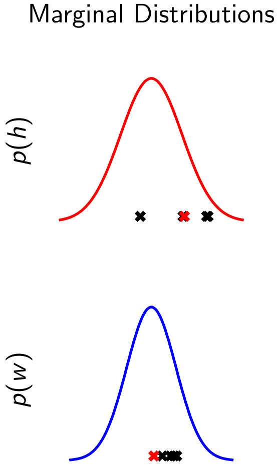
Sampling Two Dimensional Variables



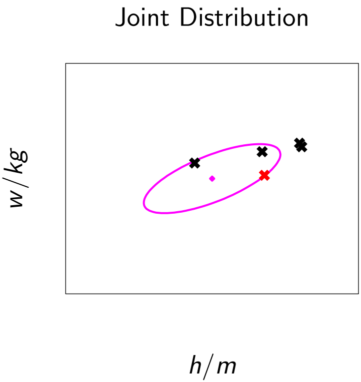


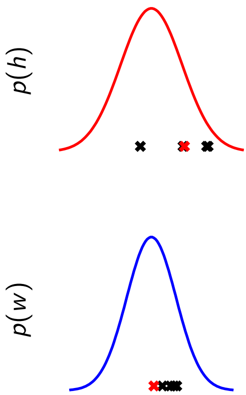
Sampling Two Dimensional Variables



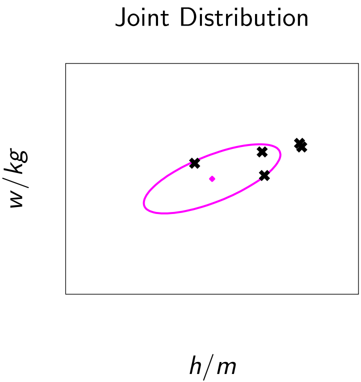


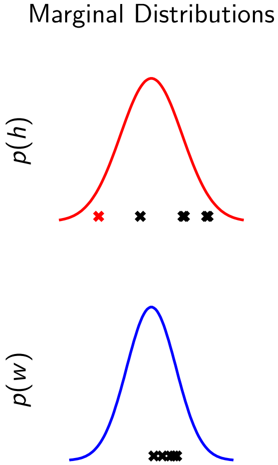
Sampling Two Dimensional Variables



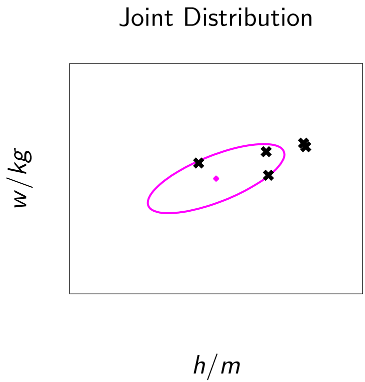


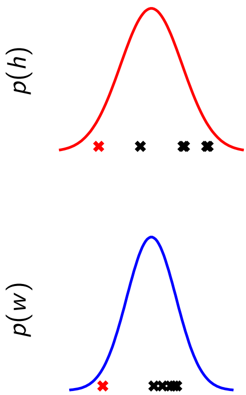
Sampling Two Dimensional Variables



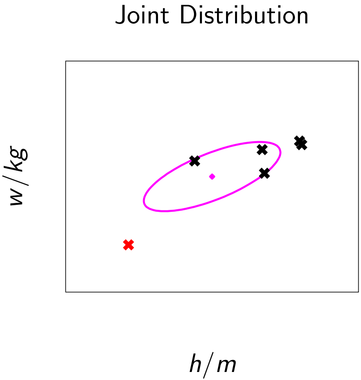


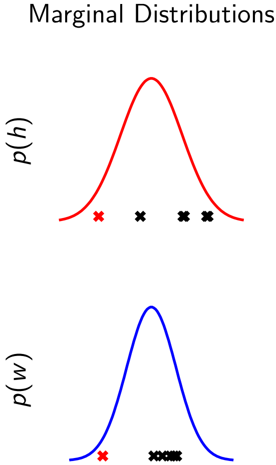
Sampling Two Dimensional Variables



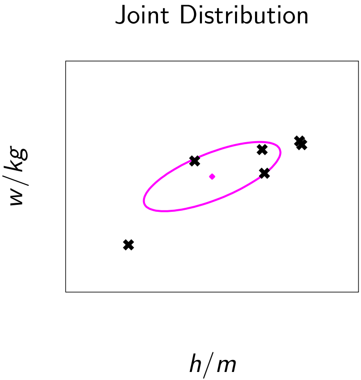


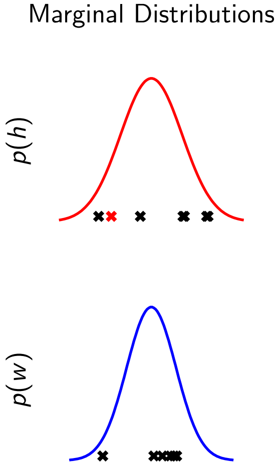
Sampling Two Dimensional Variables



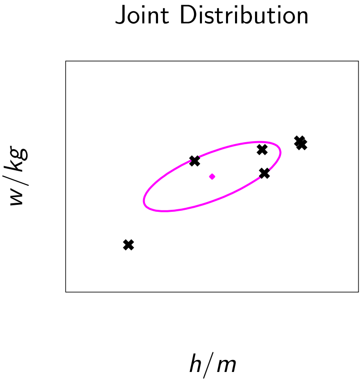


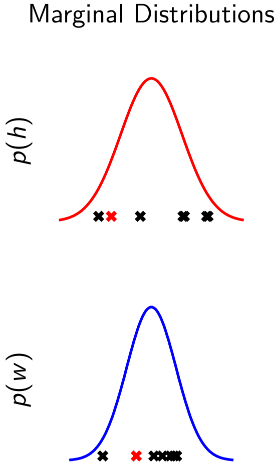
Sampling Two Dimensional Variables



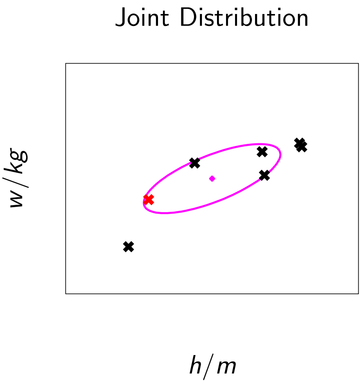


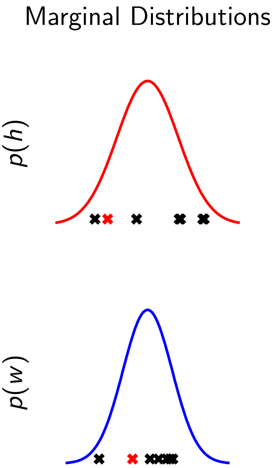
Sampling Two Dimensional Variables



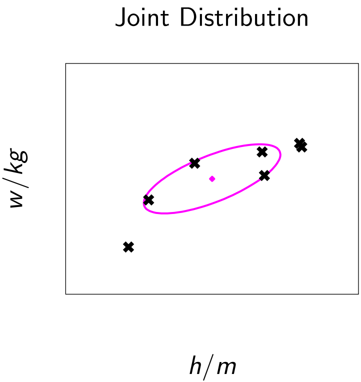


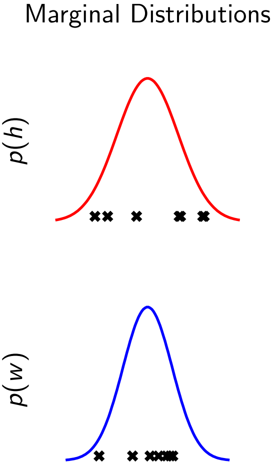
Sampling Two Dimensional Variables





Sampling Two Dimensional Variables





Independent Gaussians

$$p(w, h) = p(w)p(h)$$

Independent Gaussians

$$p(w, h) = \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{(w-\mu_1)^2}{\sigma_1^2} + \frac{(h-\mu_2)^2}{\sigma_2^2}\right)\right)$$

Independent Gaussians

$$p(w, h) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} w \\ h \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^\top \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} w \\ h \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)\right)$$

Independent Gaussians

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix **R**.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix **R**.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{R}^\top \mathbf{y} - \mathbf{R}^\top \boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{R}^\top \mathbf{y} - \mathbf{R}^\top \boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{R}\mathbf{D}^{-1}\mathbf{R}^{\top}(\mathbf{y}-\boldsymbol{\mu})\right)$$

this gives a covariance matrix:

$$\mathbf{C}^{-1} = \mathbf{R}\mathbf{D}^{-1}\mathbf{R}^{\top}$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{C}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{C}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

this gives a covariance matrix:

$$\mathbf{C} = \mathbf{R}\mathbf{D}\mathbf{R}^{\top}$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

Multivariate Consequence

- If
$$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$
- And
$$\mathbf{y} = \mathbf{W}\mathbf{x}$$
- Then
$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{W}\mu, \mathbf{W}\Sigma\mathbf{W}^\top\right)$$

Multivariate Consequence

- If
$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
- And
$$\mathbf{y} = \mathbf{W}\mathbf{x}$$
- Then
$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{W}\mu, \mathbf{W}\Sigma\mathbf{W}^{\top}\right)$$

Multivariate Consequence

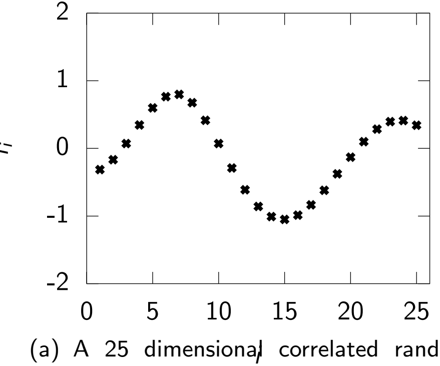
- If
$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
- And
$$\mathbf{y} = \mathbf{W}\mathbf{x}$$
- Then
$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{W}\mu, \mathbf{W}\Sigma\mathbf{W}^\top\right)$$

Sampling a Function

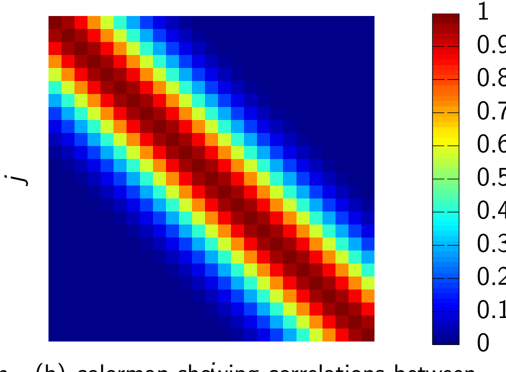
Multi-variate Gaussians

- We will consider a Gaussian with a particular structure of covariance matrix.
- Generate a single sample from this 25 dimensional Gaussian distribution, $\mathbf{f} = [f_1, f_2 \ldots f_{25}]$.
- We will plot these points against their index.

Gaussian Distribution Sample



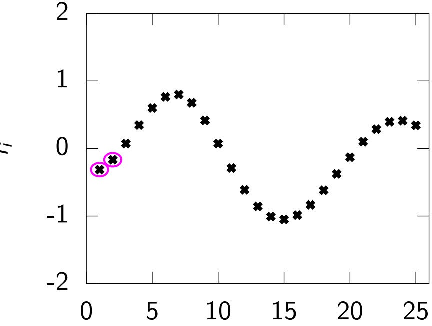
(a) A 25 dimensional correlated random variable (values ploted against index)



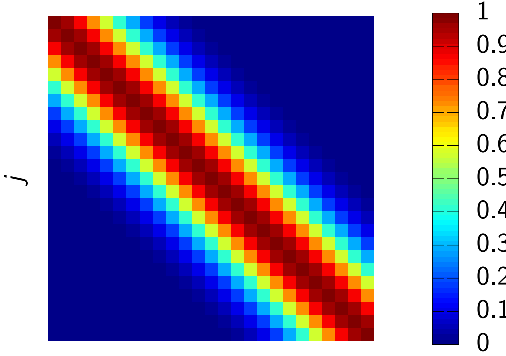
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



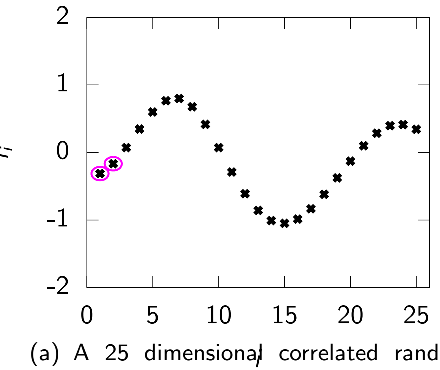
(a) A 25 dimensional correlated random variable (values ploted against index)



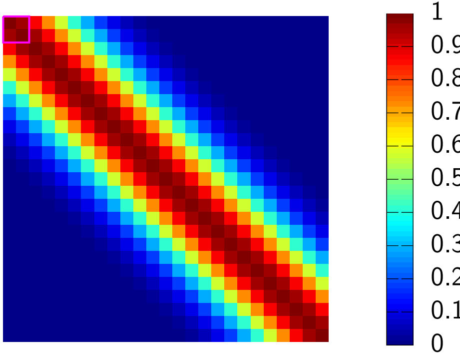
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



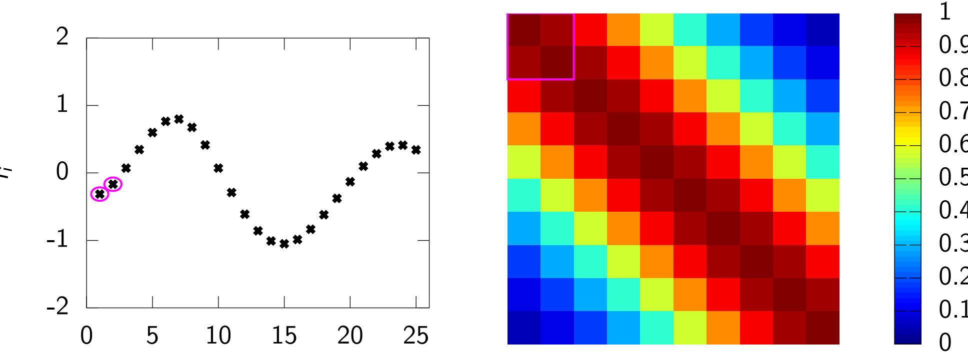
(a) A 25 dimensional correlated random variable (values ploted against index)



(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

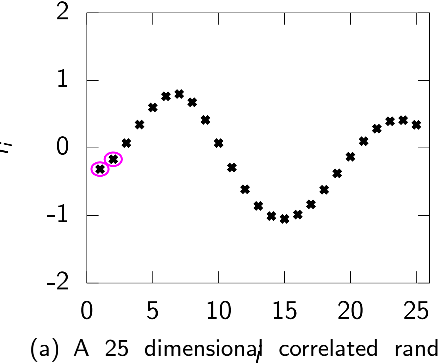


(a) A 25 dimensional correlated random variable (values ploted against index)

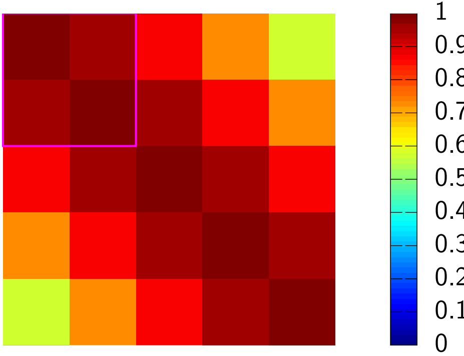
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



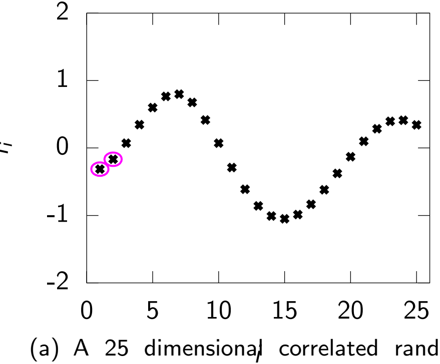
(a) A 25 dimensional correlated random variable (values ploted against index)



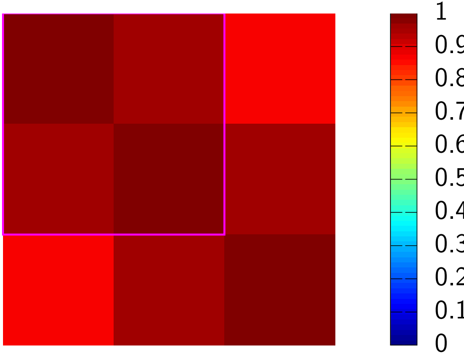
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



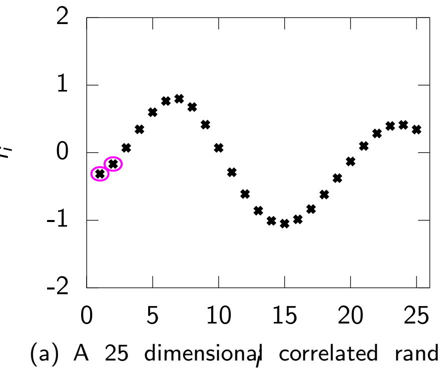
(a) A 25 dimensional correlated random variable (values ploted against index)



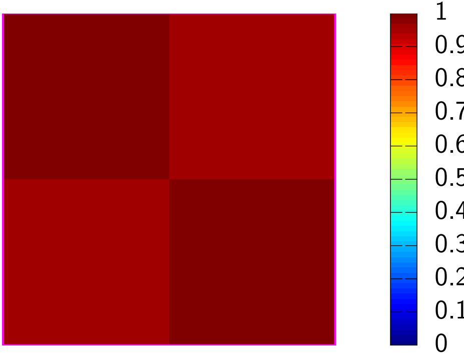
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



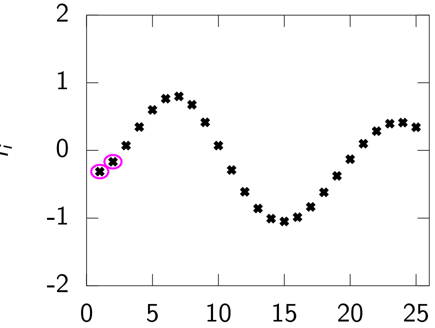
(a) A 25 dimensional correlated random variable (values ploted against index)



(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



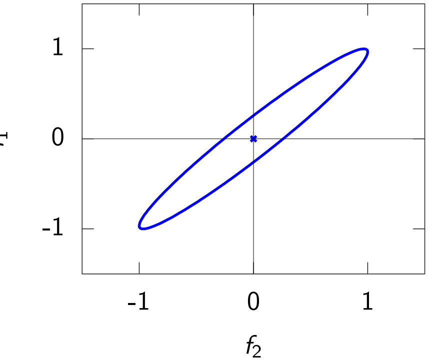
(a) A 25 dimensional correlated random variable (values ploted against index)

$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

(b) correlation between f_1 and f_2.

Figure: A sample from a 25 dimensional Gaussian distribution.

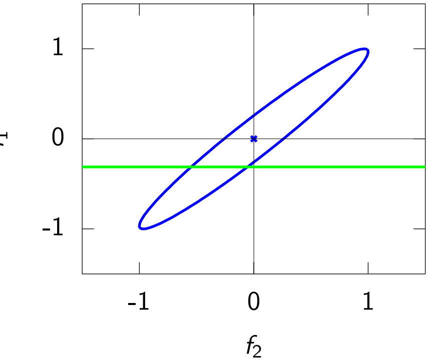
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2 | f_1 = -0.313)$.

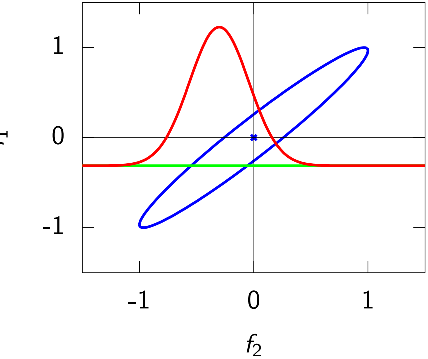
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2|f_1 = -0.313)$.

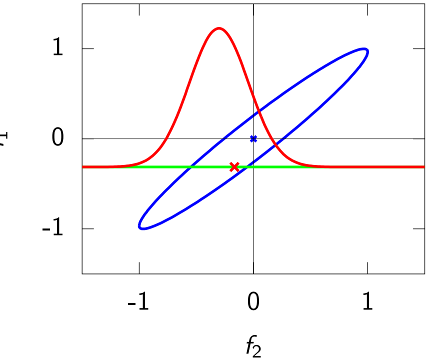
Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2 | f_1 = -0.313)$.

Prediction of f_2 from f_1



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2 | f_1 = -0.313)$.

Prediction with Correlated Gaussians

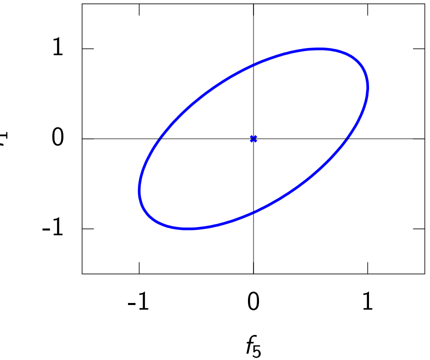
- Prediction of f_2 from f_1 requires *conditional density*.
- Conditional density is *also* Gaussian.

$$p(f_2|f_1) = \mathcal{N}\left(f_2|\frac{k_{1,2}}{k_{1,1}}f_1, k_{2,2} - \frac{k_{1,2}^2}{k_{1,1}}\right)$$

where covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} k_{1,1} & k_{1,2} \\ k_{2,1} & k_{2,2} \end{bmatrix}$$

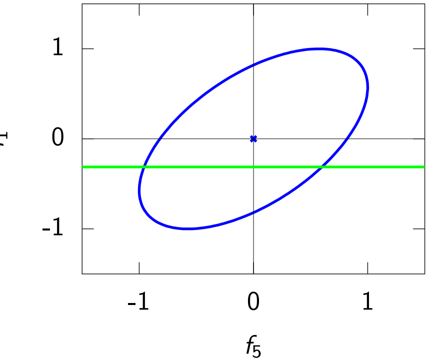
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5|f_1 = -0.313)$.

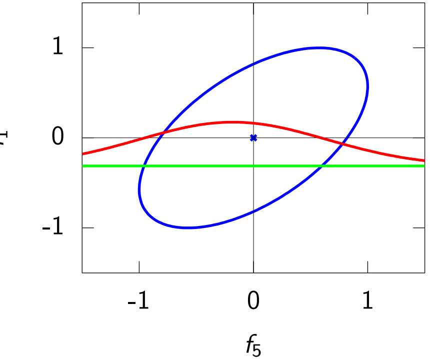
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5|f_1 = -0.313)$.

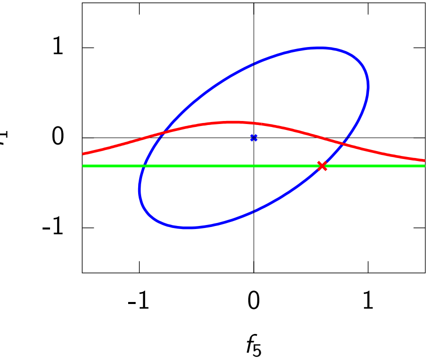
Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5|f_1 = -0.313)$.

Prediction of f_5 from f_1



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5|f_1 = -0.313)$.

Prediction with Correlated Gaussians

- Prediction of $\mathbf{f}_*$ from $\mathbf{f}$ requires multivariate *conditional density*.
- Multivariate conditional density is *also* Gaussian.

$$p(\mathbf{f}_*|\mathbf{f}) = \mathcal{N}\left(\mathbf{f}_*|\mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{f}, \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}\right)$$

- Here covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{*,\mathbf{f}} \\ \mathbf{K}_{\mathbf{f},*} & \mathbf{K}_{*,*} \end{bmatrix}$$

Prediction with Correlated Gaussians

- Prediction of $\mathbf{f}_*$ from $\mathbf{f}$ requires multivariate *conditional density*.
- Multivariate conditional density is *also* Gaussian.

$$p(\mathbf{f}_*|\mathbf{f}) = \mathcal{N}\left(\mathbf{f}_*|\boldsymbol{\mu}, \Sigma\right)$$

$$\boldsymbol{\mu} = \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{f}$$

$$\Sigma = \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}$$

- Here covariance of joint density is given by

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{*,\mathbf{f}} \\ \mathbf{K}_{\mathbf{f},*} & \mathbf{K}_{*,*} \end{bmatrix}$$

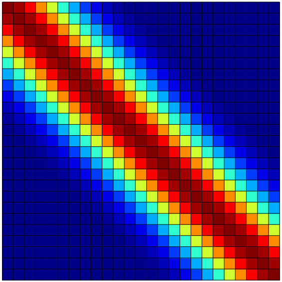
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



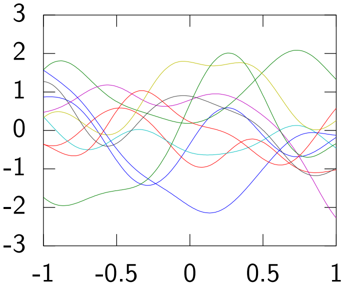
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \exp\left(-\frac{\left\|\mathbf{x} - \mathbf{x}'\right\|_2^2}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_1 = -3.0,\ x_1 = -3.0$$

$$k_{1,1} = 1.00 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 2.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_1 = -3.0$, $x_1 = -3.0$

$$k_{1,1} = 1.00 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} \boxed{1.00} & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_1 = -3.0$$

$$k_{2,1} = 1.00 \times \exp\left(-\frac{(1.20 - 1.20)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_1 = -3.0$$

$$k_{2,1} = 1.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & & \\ \boxed{0.110} & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_2 = 1.20,\ x_1 = -3.0$

$k_{2,1} = 1.00 \times \exp\left(-\frac{(1.20 - 1.20)^2}{2 \times 2.00^2}\right)$

$$\begin{bmatrix} 1.00 & 0.110 & \\ 0.110 & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_2 = 1.20$$

$$k_{2,2} = 1.00 \times \exp\left(-\frac{(1.20 - 1.20)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & \\ 0.110 & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_2 = 1.20$$

$$k_{2,2} = 1.00 \times \exp\left(-\frac{(1.20 - 1.20)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 \\ 0.110 & \boxed{1.00} \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.40,\ x_1 = -3.0$

$k_{3,1} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 2.00^2}\right)$

$$\begin{bmatrix} 1.00 & 0.110 & \\ 0.110 & 1.00 & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.40$, $x_1 = -3.0$

$$k_{3,1} = 1.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & \\ 0.110 & 1.00 & \\ \boxed{0.0889} & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_1 = -3.0$$

$$k_{3,1} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & \\ 0.0889 & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$k_{3,2} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & \\ 0.0889 & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.40,\ x_2 = 1.20$

$$k_{3,2} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & \\ 0.0889 & \boxed{0.995} & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$k_{3,2} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & 0.995 \\ 0.0889 & 0.995 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_3 = 1.40$$

$$k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & 0.995 \\ 0.0889 & 0.995 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.40,\ x_3 = 1.40$

$$k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 2.00^2}\right)$$

$$\begin{bmatrix} 1.00 & 0.110 & 0.0889 \\ 0.110 & 1.00 & 0.995 \\ 0.0889 & 0.995 & \boxed{1.00} \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

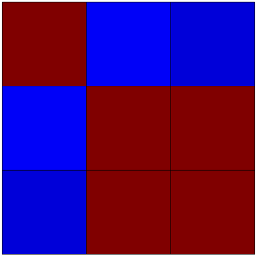
Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{\|x_i - x_j\|^2}{2\ell^2}\right)$$

$x_3 = 1.40,\ x_3 = 1.40$

$$k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 2.00^2}\right)$$



$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 2.00$ and $\alpha = 1.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_1 = -3,\ x_1 = -3$$

$$k_{1,1} = 1.0 \times \exp\left(-\frac{(-3 - -3)^2}{2 \times 2.0^2}\right)$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_1 = -3,\ x_1 = -3$

$$k_{1,1} = 1.0 \times \exp\left(-\frac{(-3 - -3)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} \boxed{1.0} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_2 = 1.2,\ x_1 = -3$$

$$k_{2,1} = 1.0 \times \exp\left(-\frac{(1.2-1.2)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_2 = 1.2,\ x_1 = -3$

$$k_{2,1} = 1.0 \times \exp\left(-\frac{(1.2-1.2)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & & & \\ \boxed{0.11} & & & \\ & & & \\ & & & \end{bmatrix}$$

$x_1 = -3,\ x_2 = 1.2,\ x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_2 = 1.2,\ x_1 = -3$$

$$k_{2,1} = 1.0 \times \exp\left(-\frac{(1.2-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & & \\ 0.11 & & & \\ & & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_2 = 1.2$, $x_2 = 1.2$

$$k_{2,2} = 1.0 \times \exp\left(-\frac{(1.2 - 1.2)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & & \\ 0.11 & & & \\ & & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_2 = 1.2,\ x_2 = 1.2$$

$$k_{2,2} = 1.0 \times \exp\left(-\frac{(1.2-1.2)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & & \\ 0.11 & \boxed{1.0} & & \\ & & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.4,\ x_1 = -3$

$$k_{3,1} = 1.0 \times \exp\left(-\frac{(1.4-1.4)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & & \\ 0.11 & 1.0 & & \\ & & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\ x_1 = -3$$

$$k_{3,1} = 1.0 \times \exp\left(-\frac{(1.4 - 1.4)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & & \\ 0.11 & 1.0 & & \\ \boxed{0.089} & & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\ x_1 = -3$$

$$k_{3,1} = 1.0 \times \exp\left(-\frac{(1.4-1.4)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & \\ 0.11 & 1.0 & & \\ 0.089 & & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.4,\ x_2 = 1.2$

$$k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4-1.4)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & \\ 0.11 & 1.0 & & \\ 0.089 & & & \\ & & & \end{bmatrix}$$

$x_1 = -3,\ x_2 = 1.2,\ x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.4,\; x_2 = 1.2$$

$$k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4 - 1.4)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & \\ 0.11 & 1.0 & & \\ 0.089 & \boxed{1.0} & & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.4,\ x_2 = 1.2$

$$k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4 - 1.4)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & & \\ & & & \end{bmatrix}$$

$x_1 = -3,\ x_2 = 1.2,\ x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.4,\ x_3 = 1.4$

$$k_{3,3} = 1.0 \times \exp\left(-\frac{(1.4 - 1.4)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & & \\ & & & \end{bmatrix}$$

$x_1 = -3,\ x_2 = 1.2,\ x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.4,\ x_3 = 1.4$

$k_{3,3} = 1.0 \times \exp\left(-\frac{(1.4-1.4)^2}{2\times 2.0^2}\right)$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & \boxed{1.0} & \\ & & & \end{bmatrix}$$

$x_1 = -3,\ x_2 = 1.2,\ x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_1 = -3$$

$$k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0 - 2.0)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & 1.0 & \\ & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_1 = -3$$

$$k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0 - 2.0)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & 1.0 & \\ \boxed{0.044} & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_4 = 2.0,\ x_1 = -3$

$$k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & & & \end{bmatrix}$$

$x_1 = -3,\ x_2 = 1.2,\ x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_2 = 1.2$$

$$k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_2 = 1.2$$

$$k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0 - 2.0)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & \boxed{0.92} & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_2 = 1.2$$

$$k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & 0.92 & & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_4 = 2.0,\ x_3 = 1.4$

$$k_{4,3} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & 0.92 & & \end{bmatrix}$$

$x_1 = -3,\ x_2 = 1.2,\ x_3 = 1.4,$ and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_3 = 1.4$$

$$k_{4,3} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2\times 2.0^2}\right)$$

$$\left[\begin{array}{cccc} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & \\ 0.044 & 0.92 & \boxed{0.96} & \end{array}\right]$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_3 = 1.4$$

$$k_{4,3} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2\times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & 0.96 \\ 0.044 & 0.92 & 0.96 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0,\ x_4 = 2.0$$

$$k_{4,4} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & 0.96 \\ 0.044 & 0.92 & 0.96 & \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_4 = 2.0, \; x_4 = 2.0$$

$$k_{4,4} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2 \times 2.0^2}\right)$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 & 0.044 \\ 0.11 & 1.0 & 1.0 & 0.92 \\ 0.089 & 1.0 & 1.0 & 0.96 \\ 0.044 & 0.92 & 0.96 & \boxed{1.0} \end{bmatrix}$$

$x_1 = -3$, $x_2 = 1.2$, $x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

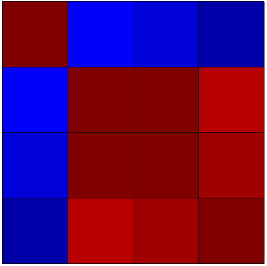
Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_4 = 2.0,\ x_4 = 2.0$

$$k_{4,4} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2 \times 2.0^2}\right)$$



$x_1 = -3,\ x_2 = 1.2,\ x_3 = 1.4$, and $x_4 = 2.0$ with $\ell = 2.0$ and $\alpha = 1.0$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_1 = -3.0,\ x_1 = -3.0$$

$$k_{1,1} = 4.00 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 5.00^2}\right)$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_1 = -3.0$, $x_1 = -3.0$

$$k_{1,1} = 4.00 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} \boxed{4.00} & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_2 = 1.20,\ x_1 = -3.0$

$k_{2,1} = 4.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2\times 5.00^2}\right)$

$$\begin{bmatrix} 4.00 & & \\ & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_2 = 1.20,\ x_1 = -3.0$

$$k_{2,1} = 4.00 \times \exp\left(-\frac{(1.20 - 1.20)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & & \\ \boxed{2.81} & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_1 = -3.0$$

$$k_{2,1} = 4.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2\times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & \\ 2.81 & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_2 = 1.20,\ x_2 = 1.20$$

$$k_{2,2} = 4.00 \times \exp\left(-\frac{(1.20 - 1.20)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & \\ 2.81 & & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_2 = 1.20,\ x_2 = 1.20$

$$k_{2,2} = 4.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2\times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & \\ 2.81 & \boxed{4.00} & \\ & & \end{bmatrix}$$

$x_1 = -3.0,\ x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\|x_i - x_j\|^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_1 = -3.0$$

$$k_{3,1} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & \\ 2.81 & 4.00 & \\ & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_1 = -3.0$$

$$k_{3,1} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & \\ 2.81 & 4.00 & \\ \boxed{2.72} & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_1 = -3.0$$

$$k_{3,1} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & \\ 2.72 & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.40,\ x_2 = 1.20$

$$k_{3,2} = 4.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & \\ 2.72 & & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.40,\ x_2 = 1.20$

$$k_{3,2} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & \\ 2.72 & \boxed{4.00} & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_2 = 1.20$$

$$k_{3,2} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & 4.00 \\ 2.72 & 4.00 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$x_3 = 1.40$, $x_3 = 1.40$

$$k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & 4.00 \\ 2.72 & 4.00 & \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$

$$x_3 = 1.40,\ x_3 = 1.40$$

$$k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right)$$

$$\begin{bmatrix} 4.00 & 2.81 & 2.72 \\ 2.81 & 4.00 & 4.00 \\ 2.72 & 4.00 & \boxed{4.00} \end{bmatrix}$$

$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

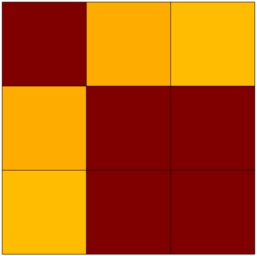
Covariance Functions

Where did this covariance matrix come from?

$$k\left(x_i, x_j\right) = \alpha \exp\left(-\frac{\|x_i - x_j\|^2}{2\ell^2}\right)$$

$x_3 = 1.40,\ x_3 = 1.40$

$$k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right)$$



$x_1 = -3.0$, $x_2 = 1.20$, and $x_3 = 1.40$ with $\ell = 5.00$ and $\alpha = 4.00$.

Outline

Basis Function Form

Radial basis functions commonly have the form

$$\phi_k(\mathbf{x}_i) = \exp\left(-\frac{\left|\mathbf{x}_i - \boldsymbol{\mu}_k\right|^2}{2\ell^2}\right).$$

- Basis function maps data into a "feature space" in which a linear sum is a non linear function.

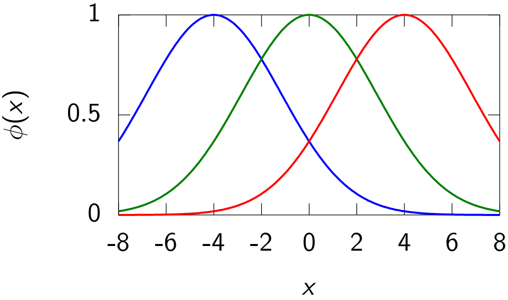


Figure: A set of radial basis functions with width $\ell = 2$ and location parameters $\boldsymbol{\mu} = [-4 \;\; 0 \;\; 4]^\top$.

Basis Function Representations

- Represent a function by a linear sum over a basis,

$$f(\mathbf{x}_{i,:}; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_{i,:}), \tag{1}$$

- Here: m basis functions and $\phi_k(\cdot)$ is kth basis function and

$$\mathbf{w} = [w_1, \ldots, w_m]^\top .$$

- For standard linear model: $\phi_k(\mathbf{x}_{i,:}) = x_{i,k}$.

Random Functions

Functions derived using:

$$f(x) = \sum_{k=1}^{m} w_k \phi_k(x),$$

where $\mathbf{W}$ is sampled from a Gaussian density,

$$w_k \sim \mathcal{N}(0, \alpha).$$

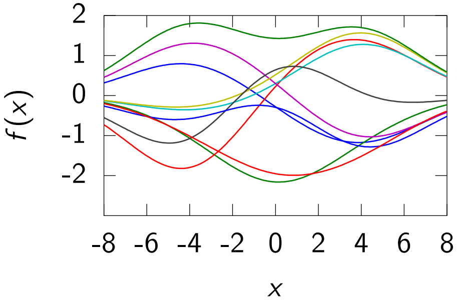


Figure: Functions sampled using the basis set from figure 2. Each line is a separate sample, generated by a weighted sum of the basis set. The weights, $\mathbf{w}$ are sampled from a Gaussian density with variance $\alpha = 1$.

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

computed at training data gives a vector

$$\mathbf{f} = \mathbf{\Phi} \mathbf{w}.$$

- $\mathbf{w}$ and $\mathbf{f}$ are only related by a inner product.
- $\mathbf{\Phi}$ is fixed and non-stochastic for a given training set.
- $\mathbf{f}$ is Gaussian distributed.
- it is straightforward to compute distribution for $\mathbf{f}$

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

computed at training data gives a vector

$$\mathbf{f} = \Phi \mathbf{w}.$$

- $\mathbf{w}$ and $\mathbf{f}$ are only related by a inner product.
- Φ is fixed and non-stochastic for a given training set.
- $\mathbf{f}$ is Gaussian distributed.
- it is straightforward to compute distribution for $\mathbf{f}$

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f\left(\mathbf{x}_i;\mathbf{w}\right)=\sum_{k=1}^{m} w_k\phi_k\left(\mathbf{x}_i\right)$$

computed at training data gives a vector

$$\mathbf{f}=\boldsymbol{\Phi}\mathbf{w}.$$

- $\mathbf{w}$ and $\mathbf{f}$ are only related by a inner product.
- $\boldsymbol{\Phi}$ is fixed and non-stochastic for a given training set.
- $\mathbf{f}$ is Gaussian distributed.
- it is straightforward to compute distribution for $\mathbf{f}$

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

computed at training data gives a vector

$$\mathbf{f} = \boldsymbol{\Phi}\mathbf{w}.$$

- **w** and **f** are only related by a inner product.
- $\boldsymbol{\Phi}$ is fixed and non-stochastic for a given training set.
- **f** is Gaussian distributed.
- it is straightforward to compute distribution for **f**

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

  computed at training data gives a vector

$$\mathbf{f} = \boldsymbol{\Phi}\mathbf{w}.$$

- $\mathbf{w}$ and $\mathbf{f}$ are only related by a inner product.
- $\boldsymbol{\Phi}$ is fixed and non-stochastic for a given training set.
- $\mathbf{f}$ is Gaussian distributed.
- it is straightforward to compute distribution for $\mathbf{f}$

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

  computed at training data gives a vector

$$\mathbf{f} = \boldsymbol{\Phi}\mathbf{w}.$$

- $\mathbf{w}$ and $\mathbf{f}$ are only related by a inner product.
- $\boldsymbol{\Phi}$ is fixed and non-stochastic for a given training set.
- $\mathbf{f}$ is Gaussian distributed.
- it is straightforward to compute distribution for $\mathbf{f}$

Direct Construction of Covariance Matrix

- Use matrix notation to write function,

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(\mathbf{x}_i)$$

  computed at training data gives a vector

$$\mathbf{f} = \mathbf{\Phi}\mathbf{w}.$$

- $\mathbf{w}$ and $\mathbf{f}$ are only related by a inner product.
- $\mathbf{\Phi}$ is fixed and non-stochastic for a given training set.
- $\mathbf{f}$ is Gaussian distributed.
- it is straightforward to compute distribution for $\mathbf{f}$

Expectations

- We use $\langle\cdot\rangle$ to denote expectations under prior distributions.
- We have

$$\langle \mathbf{f} \rangle = \phi \langle \mathbf{w} \rangle .$$

- Prior mean of $\mathbf{w}$ was zero giving

$$\langle \mathbf{f} \rangle = \mathbf{0}.$$

- Prior covariance of $\mathbf{f}$ is

$$\mathbf{K} = \left\langle \mathbf{f}\mathbf{f}^\top \right\rangle - \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^\top$$

$$\left\langle \mathbf{f}\mathbf{f}^\top \right\rangle = \boldsymbol{\Phi} \left\langle \mathbf{w}\mathbf{w}^\top \right\rangle \boldsymbol{\Phi}^\top,$$

giving

$$\mathbf{K} = \gamma' \boldsymbol{\Phi}\boldsymbol{\Phi}^\top.$$

Expectations

- We use $\langle\cdot\rangle$ to denote expectations under prior distributions.
- We have

$$\langle \mathbf{f} \rangle = \phi \langle \mathbf{w} \rangle .$$

- Prior mean of $\mathbf{w}$ was zero giving

$$\langle \mathbf{f} \rangle = \mathbf{0}.$$

- Prior covariance of $\mathbf{f}$ is

$$\mathbf{K} = \left\langle \mathbf{f}\mathbf{f}^\top \right\rangle - \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^\top$$

$$\left\langle \mathbf{f}\mathbf{f}^\top \right\rangle = \Phi \left\langle \mathbf{w}\mathbf{w}^\top \right\rangle \Phi^\top,$$

giving

$$\mathbf{K} = \gamma' \Phi \Phi^\top.$$

Expectations

- We use $\langle \cdot \rangle$ to denote expectations under prior distributions.
- We have

$$\langle \mathbf{f} \rangle = \phi \langle \mathbf{w} \rangle .$$

- Prior mean of **w** was zero giving

$$\langle \mathbf{f} \rangle = \mathbf{0}.$$

- Prior covariance of **f** is

$$\mathbf{K} = \left\langle \mathbf{f}\mathbf{f}^\top \right\rangle - \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^\top$$

$$\left\langle \mathbf{f}\mathbf{f}^\top \right\rangle = \Phi \left\langle \mathbf{w}\mathbf{w}^\top \right\rangle \Phi^\top,$$

giving

$$\mathbf{K} = \gamma' \Phi \Phi^\top.$$

Expectations

- We use $\langle \cdot \rangle$ to denote expectations under prior distributions.
- We have

$$\langle \mathbf{f} \rangle = \phi \langle \mathbf{w} \rangle .$$

- Prior mean of $\mathbf{w}$ was zero giving

$$\langle \mathbf{f} \rangle = \mathbf{0}.$$

- Prior covariance of $\mathbf{f}$ is

$$\mathbf{K} = \left\langle \mathbf{f}\mathbf{f}^{\top} \right\rangle - \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^{\top}$$

$$\left\langle \mathbf{f}\mathbf{f}^{\top} \right\rangle = \Phi \left\langle \mathbf{w}\mathbf{w}^{\top} \right\rangle \Phi^{\top},$$

giving

$$\mathbf{K} = \gamma' \Phi \Phi^{\top}.$$

Expectations

- We use $\langle\cdot\rangle$ to denote expectations under prior distributions.
- We have

$$\langle \mathbf{f} \rangle = \phi \langle \mathbf{w} \rangle .$$

- Prior mean of **w** was zero giving

$$\langle \mathbf{f} \rangle = \mathbf{0}.$$

- Prior covariance of **f** is

$$\mathbf{K} = \left\langle \mathbf{f}\mathbf{f}^\top \right\rangle - \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^\top$$

$$\left\langle \mathbf{f}\mathbf{f}^\top \right\rangle = \boldsymbol{\Phi} \left\langle \mathbf{w}\mathbf{w}^\top \right\rangle \boldsymbol{\Phi}^\top,$$

giving

$$\mathbf{K} = \gamma' \boldsymbol{\Phi}\boldsymbol{\Phi}^\top.$$

Covariance between Two Points

- The prior covariance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$k(\mathbf{x}_i, \mathbf{x}_j) = \gamma' \sum_{\ell}^{m} \phi_\ell(\mathbf{x}_i)\, \phi_\ell(\mathbf{x}_j)$$

or in vector form

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi_:(\mathbf{x}_i)^\top \phi_:(\mathbf{x}_j),$$

- For the radial basis used this gives

$$k(\mathbf{x}_i, \mathbf{x}_j) = \gamma' \sum_{k=1}^{m} \exp\left(-\frac{|\mathbf{x}_i - \boldsymbol{\mu}_k|^2 + |\mathbf{x}_j - \boldsymbol{\mu}_k|^2}{2\ell^2}\right).$$

Covariance between Two Points

- The prior covariance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \gamma' \sum_{\ell}^{m} \phi_\ell\left(\mathbf{x}_i\right) \phi_\ell\left(\mathbf{x}_j\right)$$

  or in vector form

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi_:\left(\mathbf{x}_i\right)^\top \phi_:\left(\mathbf{x}_j\right),$$

- For the radial basis used this gives

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \gamma' \sum_{k=1}^{m} \exp\left(-\frac{\left|\mathbf{x}_i - \boldsymbol{\mu}_k\right|^2 + \left|\mathbf{x}_j - \boldsymbol{\mu}_k\right|^2}{2\ell^2}\right).$$

Covariance between Two Points

- The prior covariance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \gamma' \sum_{\ell}^{m} \phi_\ell\left(\mathbf{x}_i\right) \phi_\ell\left(\mathbf{x}_j\right)$$

  or in vector form

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi_:\left(\mathbf{x}_i\right)^\top \phi_:\left(\mathbf{x}_j\right),$$

- For the radial basis used this gives

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \gamma' \sum_{k=1}^{m} \exp\left(-\frac{\left|\mathbf{x}_i - \boldsymbol{\mu}_k\right|^2 + \left|\mathbf{x}_j - \boldsymbol{\mu}_k\right|^2}{2\ell^2}\right).$$

Covariance between Two Points

- The prior covariance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \gamma' \sum_{\ell}^{m} \phi_\ell\left(\mathbf{x}_i\right) \phi_\ell\left(\mathbf{x}_j\right)$$

  or in vector form

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi_:\left(\mathbf{x}_i\right)^\top \phi_:\left(\mathbf{x}_j\right),$$

- For the radial basis used this gives

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \gamma' \sum_{k=1}^{m} \exp\left(-\frac{\left|\mathbf{x}_i - \boldsymbol{\mu}_k\right|^2 + \left|\mathbf{x}_j - \boldsymbol{\mu}_k\right|^2}{2\ell^2}\right).$$

Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions
- Consider uniform spacing over a region:

$$k\left(x_i, x_j\right) = \gamma \Delta\mu \sum_{k=1}^{m} \exp\left(-\frac{x_i^2 + x_j^2 - 2\mu_k\left(x_i + x_j\right) + 2\mu_k^2}{2\ell^2}\right),$$

Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions
- Consider uniform spacing over a region:

$$k\left(x_i, x_j\right) = \gamma \Delta\mu \sum_{k=1}^{m} \exp\left(-\frac{x_i^2 + x_j^2 - 2\mu_k\left(x_i + x_j\right) + 2\mu_k^2}{2\ell^2}\right),$$

Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions
- Consider uniform spacing over a region:

$$k(x_i, x_j) = \gamma \Delta\mu \sum_{k=1}^{m} \exp\left(-\frac{x_i^2 + x_j^2 - 2\mu_k(x_i + x_j) + 2\mu_k^2}{2\ell^2}\right),$$

Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions
- Consider uniform spacing over a region:

$$k\left(x_i, x_j\right) = \gamma \Delta\mu \sum_{k=1}^{m} \exp\left(-\frac{x_i^2 + x_j^2 - 2\mu_k\left(x_i + x_j\right) + 2\mu_k^2}{2\ell^2}\right),$$

Uniform Basis Functions

- Set each center location to

$$\mu_k = a + \Delta\mu \cdot (k - 1).$$

- Specify the bases in terms of their indices,

$$k(x_i, x_j) = \gamma\Delta\mu \sum_{k=1}^{m} \exp\left(-\frac{x_i^2 + x_j^2}{2\ell^2} - \frac{2(a + \Delta\mu \cdot k)(x_i + x_j) + 2(a + \Delta\mu \cdot k)^2}{2\ell^2} \right).$$

Uniform Basis Functions

- Set each center location to

$$\mu_k = a + \Delta\mu \cdot (k - 1).$$

- Specify the bases in terms of their indices,

$$\begin{aligned} k\left(x_i, x_j\right) =& \gamma \Delta\mu \sum_{k=1}^{m} \exp\Bigg(- \frac{x_i^2 + x_j^2}{2\ell^2} \\ & - \frac{2\left(a + \Delta\mu \cdot k\right)\left(x_i + x_j\right) + 2\left(a + \Delta\mu \cdot k\right)^2}{2\ell^2} \Bigg). \end{aligned}$$

Infinite Basis Functions

- Take $\mu_0 = a$ and $\mu_m = b$ so $b = a + \Delta\mu \cdot (m-1)$.
- Take limit as $\Delta\mu \to 0$ so $m \to \infty$

$$
\begin{aligned}
k(x_i, x_j) =& \gamma \int_a^b \exp\Bigg(-\frac{x_i^2 + x_j^2}{2\ell^2} \\
&+ \frac{2\left(\mu - \frac{1}{2}(x_i + x_j)\right)^2 - \frac{1}{2}(x_i + x_j)^2}{2\ell^2} \Bigg) \mathrm{d}\mu,
\end{aligned}
$$

where we have used $k \cdot \Delta\mu \to \mu$.

Infinite Basis Functions

- Take $\mu_0 = a$ and $\mu_m = b$ so $b = a + \Delta\mu \cdot (m-1)$.
- Take limit as $\Delta\mu \to 0$ so $m \to \infty$

$$k(x_i, x_j) = \gamma \int_a^b \exp\left(-\frac{x_i^2 + x_j^2}{2\ell^2} + \frac{2\left(\mu - \frac{1}{2}\left(x_i + x_j\right)\right)^2 - \frac{1}{2}\left(x_i + x_j\right)^2}{2\ell^2} \right) \mathrm{d}\mu,$$

where we have used $k \cdot \Delta\mu \to \mu$.

Infinite Basis Functions

- Take $\mu_0 = a$ and $\mu_m = b$ so $b = a + \Delta\mu \cdot (m - 1)$.
- Take limit as $\Delta\mu \to 0$ so $m \to \infty$

$$k(x_i, x_j) = \gamma \int_a^b \exp\left(-\frac{x_i^2 + x_j^2}{2\ell^2} + \frac{2\left(\mu - \frac{1}{2}(x_i + x_j)\right)^2 - \frac{1}{2}(x_i + x_j)^2}{2\ell^2} \right) \mathrm{d}\mu,$$

where we have used $k \cdot \Delta\mu \to \mu$.

Infinite Basis Functions

- Take $\mu_0 = a$ and $\mu_m = b$ so $b = a + \Delta\mu \cdot (m-1)$.
- Take limit as $\Delta\mu \to 0$ so $m \to \infty$

$$k(x_i, x_j) = \gamma \int_a^b \exp\Bigg(- \frac{x_i^2 + x_j^2}{2\ell^2} + \frac{2\left(\mu - \frac{1}{2}\left(x_i + x_j\right)\right)^2 - \frac{1}{2}\left(x_i + x_j\right)^2}{2\ell^2} \Bigg) \mathrm{d}\mu,$$

where we have used $k \cdot \Delta\mu \to \mu$.

Result

- Performing the integration leads to

$$k(x_i,x_j) = \gamma \frac{\sqrt{\pi \ell^2}}{2} \exp\left(-\frac{(x_i - x_j)^2}{4\ell^2}\right) \times \left[\operatorname{erf}\left(\frac{\left(b - \frac{1}{2}(x_i + x_j)\right)}{\ell}\right) - \operatorname{erf}\left(\frac{\left(a - \frac{1}{2}(x_i + x_j)\right)}{\ell}\right)\right],$$

- Now take limit as $a \to -\infty$ and $b \to \infty$

$$k(x_i, x_j) = \alpha \exp\left(-\frac{(x_i - x_j)^2}{4\ell^2}\right).$$

where $\alpha = \gamma\sqrt{\pi\ell^2}$.

Result

- Performing the integration leads to

$$k(x_i, x_j) = \gamma \frac{\sqrt{\pi \ell^2}}{2} \exp\left(-\frac{(x_i - x_j)^2}{4\ell^2}\right) \times \left[\operatorname{erf}\left(\frac{\left(b - \frac{1}{2}(x_i + x_j)\right)}{\ell}\right) - \operatorname{erf}\left(\frac{\left(a - \frac{1}{2}(x_i + x_j)\right)}{\ell}\right)\right],$$

- Now take limit as $a \to -\infty$ and $b \to \infty$

$$k(x_i, x_j) = \alpha \exp\left(-\frac{(x_i - x_j)^2}{4\ell^2}\right).$$

where $\alpha = \gamma\sqrt{\pi \ell^2}$.

Result

- Performing the integration leads to

$$k(x_i,x_j) = \gamma \frac{\sqrt{\pi\ell^2}}{2} \exp\left(-\frac{(x_i - x_j)^2}{4\ell^2}\right)$$
$$\times \left[\operatorname{erf}\left(\frac{\left(b - \frac{1}{2}(x_i + x_j)\right)}{\ell}\right) - \operatorname{erf}\left(\frac{\left(a - \frac{1}{2}(x_i + x_j)\right)}{\ell}\right)\right],$$

- Now take limit as $a \to -\infty$ and $b \to \infty$

$$k(x_i, x_j) = \alpha \exp\left(-\frac{(x_i - x_j)^2}{4\ell^2}\right).$$

where $\alpha = \gamma\sqrt{\pi\ell^2}$.

Infinite Feature Space

- A RBF model with infinite basis functions is a Gaussian process.
- The covariance function is the exponentiated quadratic.
- **Note:** The functional form for the covariance function and basis functions are similar.
  - this is a special case,
  - in general they are very different
- Similar results can obtained for multi-dimensional input networks **??**.

Infinite Feature Space

- A RBF model with infinite basis functions is a Gaussian process.
- The covariance function is the exponentiated quadratic.
- **Note:** The functional form for the covariance function and basis functions are similar.
  - this is a special case,
  - in general they are very different
- Similar results can obtained for multi-dimensional input networks **??**.

Infinite Feature Space

- A RBF model with infinite basis functions is a Gaussian process.
- The covariance function is the exponentiated quadratic.
- **Note:** The functional form for the covariance function and basis functions are similar.
  - this is a special case,
  - in general they are very different
- Similar results can obtained for multi-dimensional input networks **??**.

Infinite Feature Space

- A RBF model with infinite basis functions is a Gaussian process.
- The covariance function is the exponentiated quadratic.
- **Note:** The functional form for the covariance function and basis functions are similar.
  - this is a special case,
  - in general they are very different
- Similar results can obtained for multi-dimensional input networks **??**.

Nonparametric Gaussian Processes

- This work takes us from parametric to non-parametric.
- The limit implies infinite dimensional $\mathbf{w}$.
- Gaussian processes are generally non-parametric: combine data with covariance function to get model.
- This representation *cannot* be summarized by a parameter vector of a fixed size.

The Parametric Bottleneck

- Parametric models have a representation that does not respond to increasing training set size.
- Bayesian posterior distributions over parameters contain the information about the training data.
  - Use Bayes' rule from training data, $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$,
  - Make predictions on test data

$$p(y_*|\mathbf{X}_*, \mathbf{y}, \mathbf{X}) = \int p(y_*|\mathbf{w}, \mathbf{X}_*)\, p(\mathbf{w}|\mathbf{y}, \mathbf{X})\mathrm{d}\mathbf{w})\,.$$

- $\mathbf{w}$ becomes a bottleneck for information about the training set to pass to the test set.
- Solution: increase m so that the bottleneck is so large that it no longer presents a problem.
- How big is big enough for m? Non-parametrics says $m \to \infty$.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form given in (1).
- However, it *is* possible to express *parametric* as GPs:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi_:(\mathbf{x}_i)^\top \phi_:(\mathbf{x}_j).$$

- These are known as degenerate covariance matrices.
- Their rank is at most m, non-parametric models have full rank covariance matrices.
- Most well known is the "linear kernel", $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form given in (1).
- However, it *is* possible to express *parametric* as GPs:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi_:(\mathbf{x}_i)^\top \phi_:(\mathbf{x}_j).$$

- These are known as degenerate covariance matrices.
- Their rank is at most m, non-parametric models have full rank covariance matrices.
- Most well known is the "linear kernel", $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form given in (1).
- However, it *is* possible to express *parametric* as GPs:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi_{:}(\mathbf{x}_i)^\top \phi_{:}(\mathbf{x}_j).$$

- These are known as degenerate covariance matrices.
- Their rank is at most m, non-parametric models have full rank covariance matrices.
- Most well known is the "linear kernel", $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form given in (1).
- However, it *is* possible to express *parametric* as GPs:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi_:(\mathbf{x}_i)^\top \phi_:(\mathbf{x}_j).$$

- These are known as degenerate covariance matrices.
- Their rank is at most m, non-parametric models have full rank covariance matrices.
- Most well known is the "linear kernel", $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form given in (1).
- However, it *is* possible to express *parametric* as GPs:

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi_:\left(\mathbf{x}_i\right)^\top \phi_:\left(\mathbf{x}_j\right).$$

- These are known as degenerate covariance matrices.
- Their rank is at most m, non-parametric models have full rank covariance matrices.
- Most well known is the "linear kernel", $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a *fixed* number of parameters.
- Complexity of parametric model remains fixed regardless of the size of our training data set.
- For a non-parametric model the required number of parameters grows with the size of the training data.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a *fixed* number of parameters.
- Complexity of parametric model remains fixed regardless of the size of our training data set.
- For a non-parametric model the required number of parameters grows with the size of the training data.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a *fixed* number of parameters.
- Complexity of parametric model remains fixed regardless of the size of our training data set.
- For a non-parametric model the required number of parameters grows with the size of the training data.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a *fixed* number of parameters.
- Complexity of parametric model remains fixed regardless of the size of our training data set.
- For a non-parametric model the required number of parameters grows with the size of the training data.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a *fixed* number of parameters.
- Complexity of parametric model remains fixed regardless of the size of our training data set.
- For a non-parametric model the required number of parameters grows with the size of the training data.

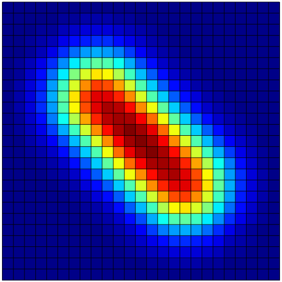
Covariance Functions

RBF Basis Functions

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \boldsymbol{\phi}(\mathbf{x})^{\top} \boldsymbol{\phi}(\mathbf{x}')$$

$$\phi_i(x) = \exp\left(-\frac{\|x - \mu_i\|_2^2}{\ell^2}\right)$$

$$\boldsymbol{\mu} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$



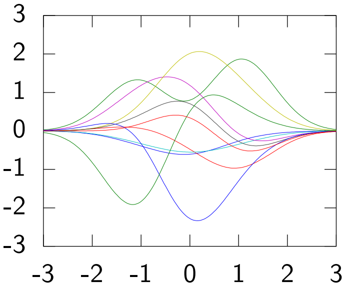
Covariance Functions

RBF Basis Functions

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}')$$

$$\phi_i(x) = \exp\left(-\frac{\|x - \mu_i\|_2^2}{\ell^2}\right)$$

$$\boldsymbol{\mu} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$



Covariance Functions and Mercer Kernels

- Mercer Kernels and Covariance Functions are similar.
- the kernel perspective does not make a probabilistic interpretation of the covariance function.
- Algorithms can be simpler, but probabilistic interpretation is crucial for kernel parameter optimization.

Covariance Functions and Mercer Kernels

- Mercer Kernels and Covariance Functions are similar.
- the kernel perspective does not make a probabilistic interpretation of the covariance function.
- Algorithms can be simpler, but probabilistic interpretation is crucial for kernel parameter optimization.

Covariance Functions and Mercer Kernels

- Mercer Kernels and Covariance Functions are similar.
- the kernel perspective does not make a probabilistic interpretation of the covariance function.
- Algorithms can be simpler, but probabilistic interpretation is crucial for kernel parameter optimization.

Outline

Constructing Covariance Functions

- Sum of two covariances is also a covariance function.

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$

Constructing Covariance Functions

- Product of two covariances is also a covariance function.

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$$

Multiply by Deterministic Function

- If $f(\mathbf{x})$ is a Gaussian process.
- $g(\mathbf{x})$ is a deterministic function.
- $h(\mathbf{x}) = f(\mathbf{x})g(\mathbf{x})$
- Then

$$k_h(\mathbf{x}, \mathbf{x}') = g(\mathbf{x})k_f(\mathbf{x}, \mathbf{x}')g(\mathbf{x}')$$

  where k_h is covariance for $h(\cdot)$ and k_f is covariance for $f(\cdot)$.

Covariance Functions

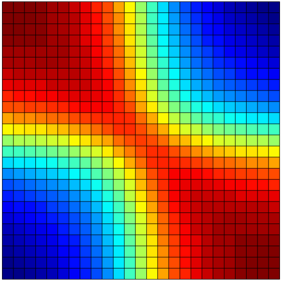
MLP Covariance Function

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \mathrm{asin}\left(\frac{w\mathbf{x}^{\top}\mathbf{x}' + b}{\sqrt{w\mathbf{x}^{\top}\mathbf{x} + b + 1}\sqrt{w\mathbf{x}'^{\top}\mathbf{x}' + b + 1}}\right)$$

- Based on infinite neural network model.

$$w = 40$$

$$b = 4$$



Covariance Functions

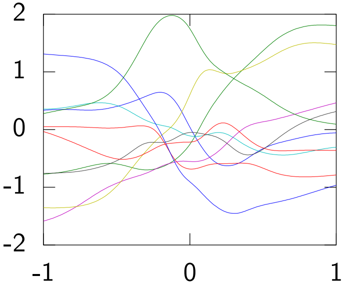
MLP Covariance Function

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \text{asin}\left(\frac{w\mathbf{x}^\top \mathbf{x}' + b}{\sqrt{w\mathbf{x}^\top \mathbf{x} + b + 1}\sqrt{w\mathbf{x}'^\top \mathbf{x}' + b + 1}}\right)$$

- Based on infinite neural network model.

  $$w = 40$$

  $$b = 4$$



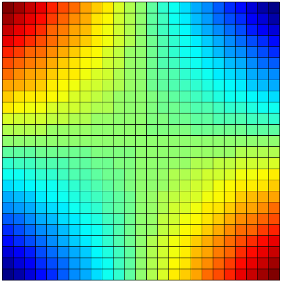
Covariance Functions

Linear Covariance Function

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \mathbf{x}^\top \mathbf{x}'$$

- Bayesian linear regression.

$$\alpha = 1$$



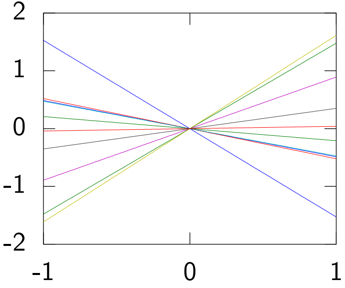
Covariance Functions

Linear Covariance Function

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \mathbf{x}^\top \mathbf{x}'$$

- Bayesian linear regression.

$$\alpha = 1$$



Gaussian Process Interpolation

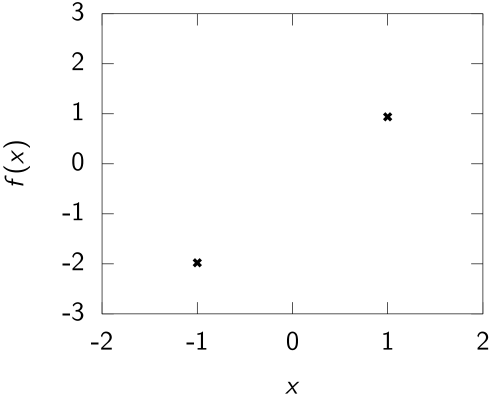


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

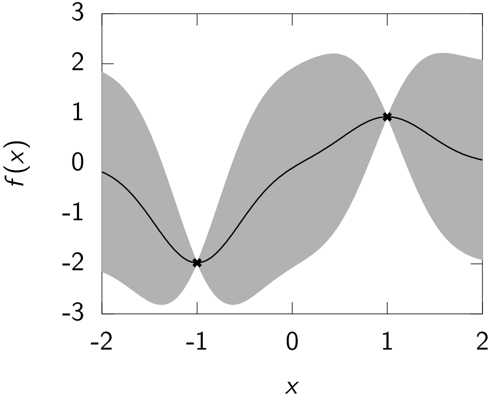


Figure: Real example: BACCO (see *e.g.* (**?**)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

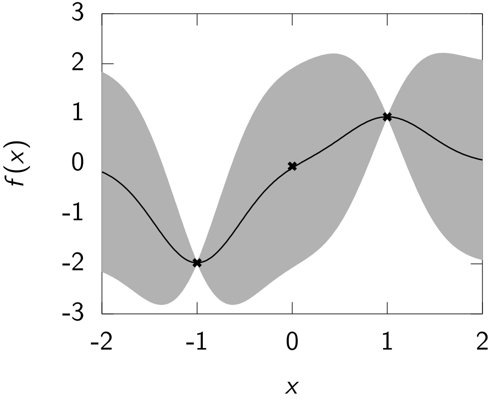


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

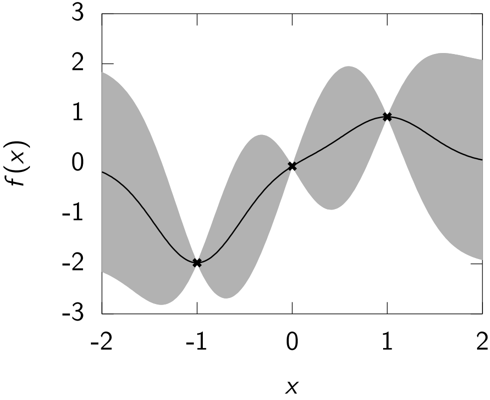


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

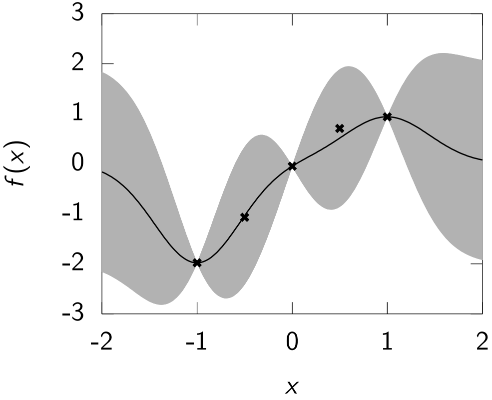


Figure: Real example: BACCO (see *e.g.* (**?**)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

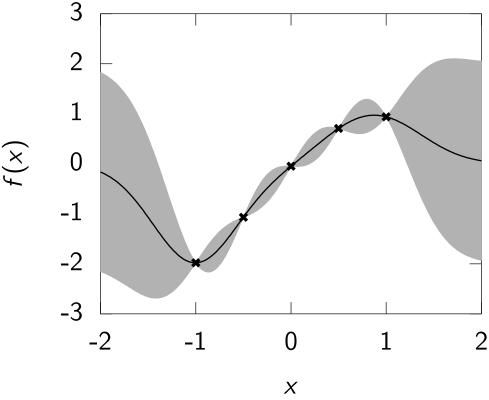


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

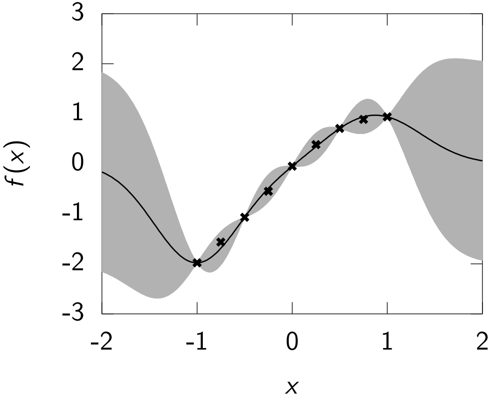


Figure: Real example: BACCO (see *e.g.* (?)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Gaussian Process Interpolation

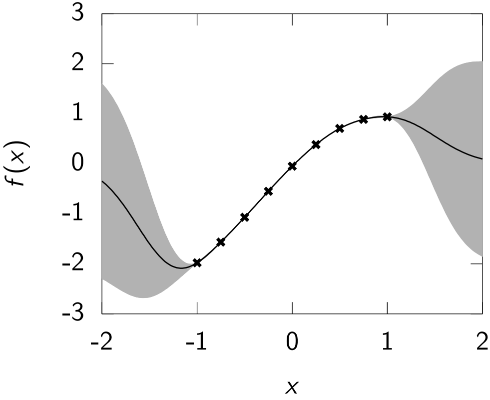


Figure: Real example: BACCO (see *e.g.* (**?**)). Interpolation through outputs from slow computer simulations (*e.g.* atmospheric carbon levels).

Noise Models

Graph of a GP

- Relates input variables, $\mathbf{X}$, to vector, $\mathbf{y}$, through $\mathbf{f}$ given kernel parameters $\boldsymbol{\theta}$.
- Plate notation indicates independence of $y_i|f_i$.
- Noise model, $p(y_i|f_i)$ can take several forms.
- Simplest is Gaussian noise.

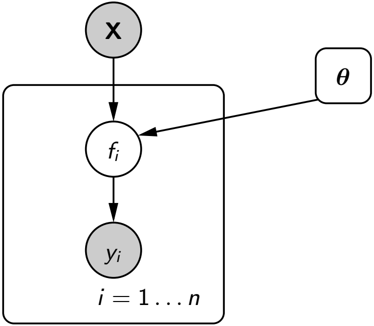


Figure: The Gaussian process depicted graphically.

Gaussian Noise

- Gaussian noise model,

$$p\left(y_i|f_i\right) = \mathcal{N}\left(y_i|f_i, \sigma^2\right)$$

  where σ^2 is the variance of the noise.
- Equivalent to a covariance function of the form

$$k(\mathbf{x}_i, \mathbf{x}_j) = \delta_{i,j}\sigma^2$$

  where $\delta_{i,j}$ is the Kronecker delta function.
- Additive nature of Gaussians means we can simply add this term to existing covariance matrices.

Gaussian Process Regression

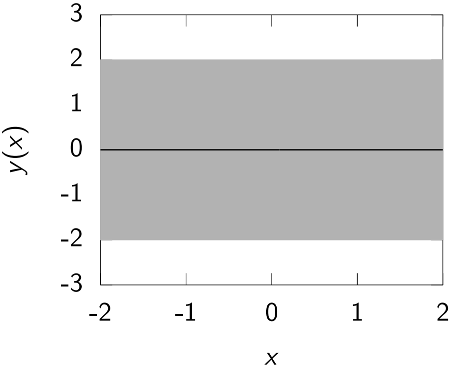


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

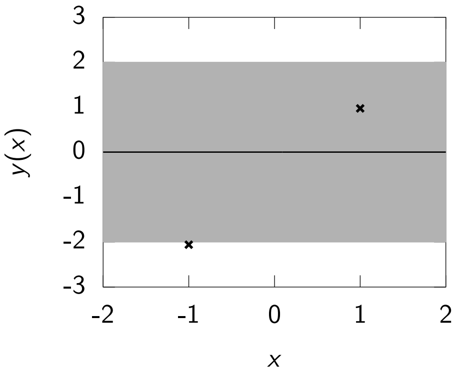


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

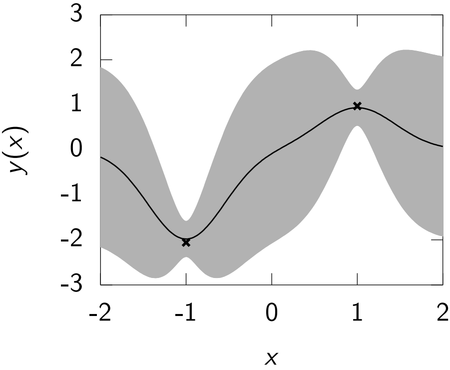


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

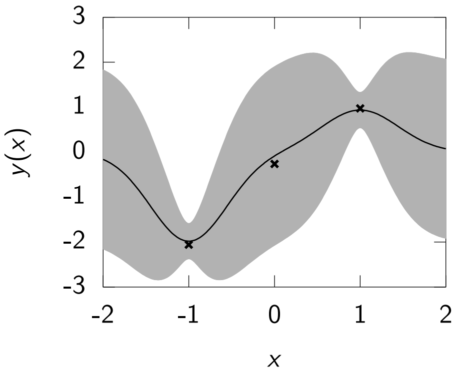


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

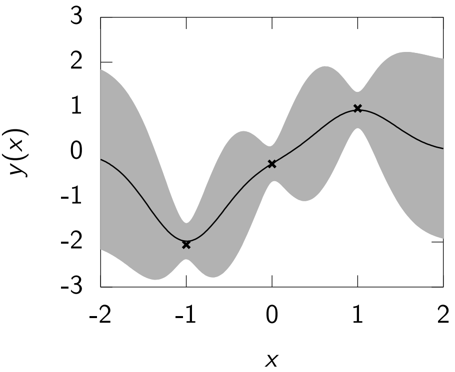


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

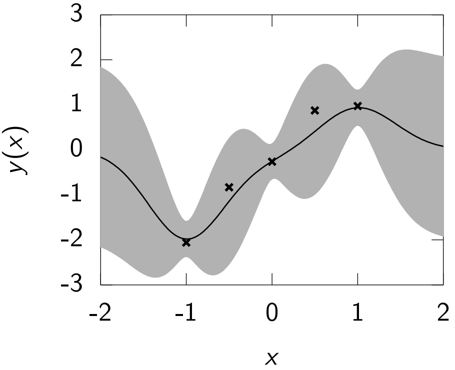


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

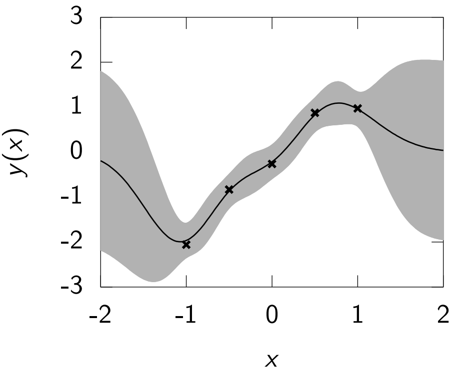


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

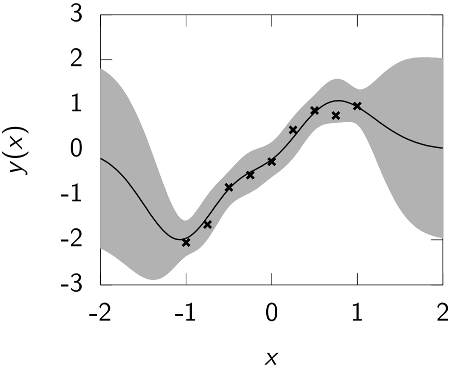


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

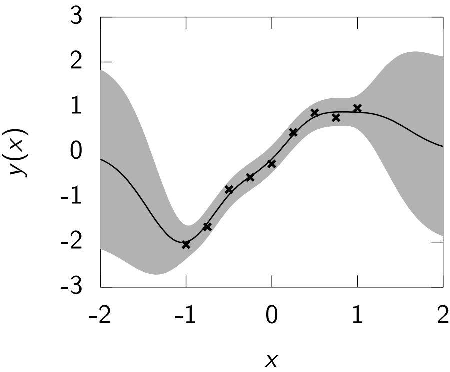


Figure: Examples include WiFi localization, C14 callibration curve.

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

$$\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{K}|}\exp\left(-\frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}\right)$$

The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

$$\mathcal{N}\left(\mathbf{y}|\mathbf{0},\mathbf{K}\right) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{K}|}\exp\left(-\frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}\right)$$

The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

$$\log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}\right) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

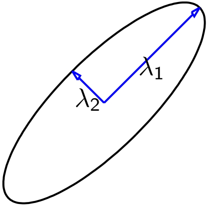
$$E(\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{K}| + \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$$

Eigendecomposition of Covariance

$$\mathbf{K} = \mathbf{R}\mathbf{\Lambda}^2\mathbf{R}^\top$$

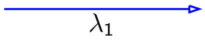


where $\mathbf{\Lambda}$ is a *diagonal* matrix and $\mathbf{R}^\top\mathbf{R} = \mathbf{I}$.

Useful representation since $|\mathbf{K}| = \left|\mathbf{\Lambda}^2\right| = |\mathbf{\Lambda}|^2$.

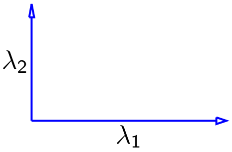
Capacity control: $\log |\mathbf{K}|$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



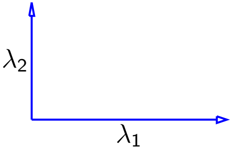
Capacity control: $\log |\mathbf{K}|$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



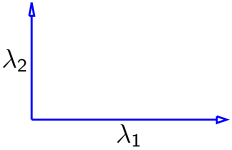
Capacity control: $\log |\mathbf{K}|$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



Capacity control: $\log |\mathbf{K}|$

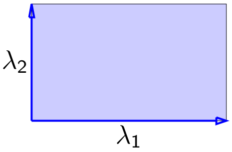
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

Capacity control: $\log |\mathbf{K}|$

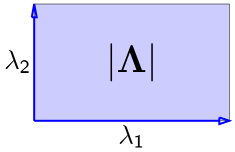
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

Capacity control: $\log |\mathbf{K}|$

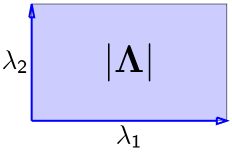
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

Capacity control: $\log |\mathbf{K}|$

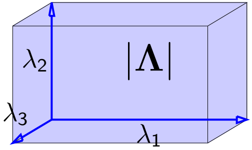
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

Capacity control: $\log |\mathbf{K}|$

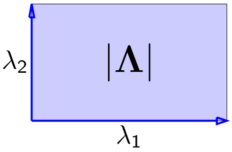
$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ \boxed{0 \quad 0 \quad \lambda_3} \end{bmatrix}$$



$$|\boldsymbol{\Lambda}| = \lambda_1 \lambda_2 \lambda_3$$

Capacity control: $\log |\mathbf{K}|$

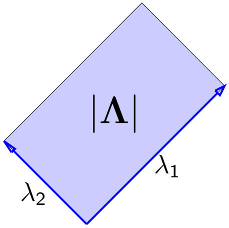
$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$



$$|\mathbf{\Lambda}| = \lambda_1 \lambda_2$$

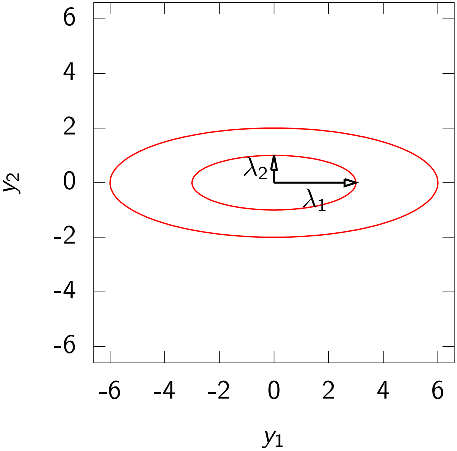
Capacity control: $\log |\mathbf{K}|$

$$\mathbf{R}\boldsymbol{\Lambda} = \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix}$$

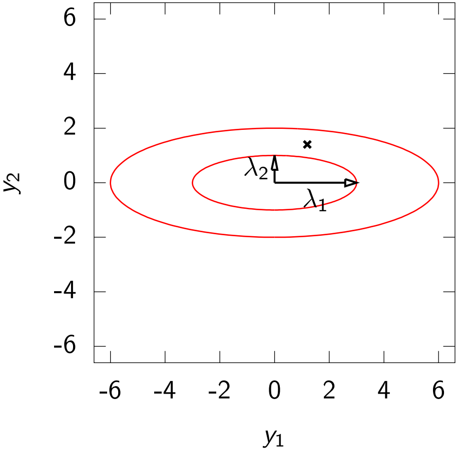


$$|\mathbf{R}\boldsymbol{\Lambda}| = \lambda_1 \lambda_2$$

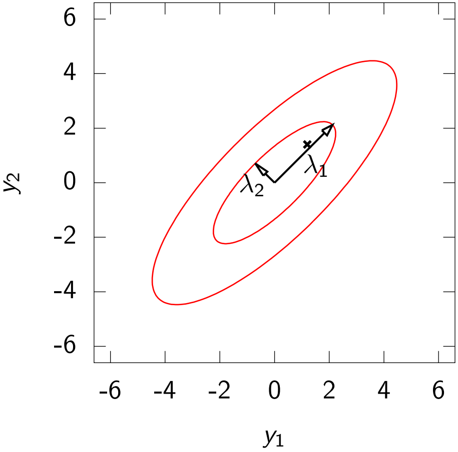
Data Fit: $\frac{\mathbf{y}^{-1}\mathbf{K}^{-1}\mathbf{y}}{2}$



Data Fit: $\frac{\mathbf{y}^{-1}\mathbf{K}^{-1}\mathbf{y}}{2}$

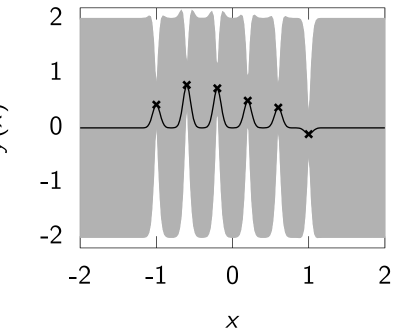


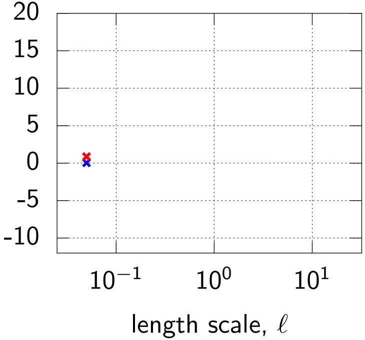
Data Fit: $\frac{\mathbf{y}^{-1}\mathbf{K}^{-1}\mathbf{y}}{2}$



Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

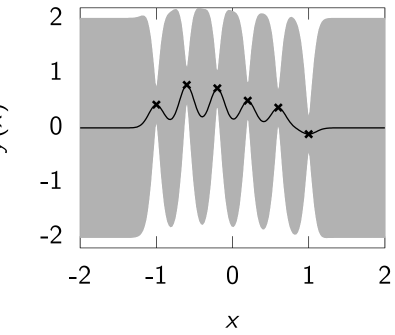


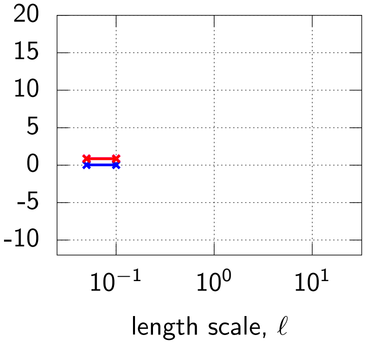


$$E(\boldsymbol{\theta}) = \frac{1}{2}\left|\mathbf{K}\right| + \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

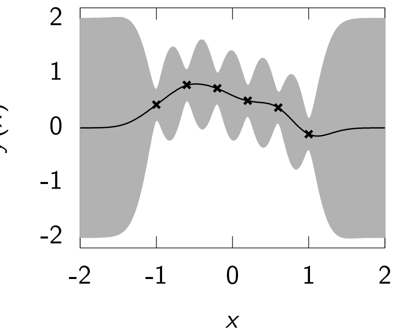


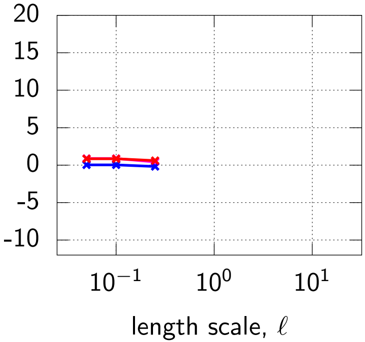


$$E(\boldsymbol{\theta}) = \frac{1}{2}|\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

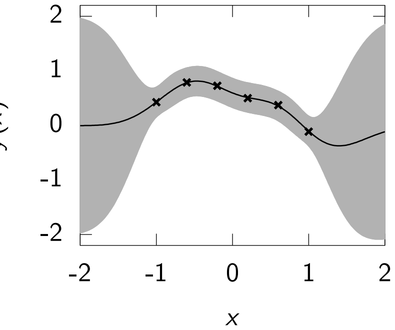


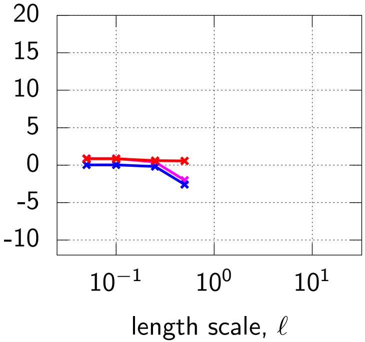


$$E(\boldsymbol{\theta}) = \frac{1}{2}\left|\mathbf{K}\right| + \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

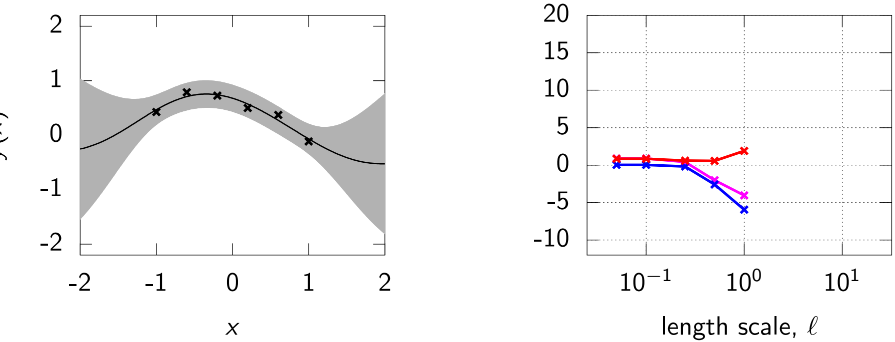




$$E(\boldsymbol{\theta}) = \frac{1}{2}\left|\mathbf{K}\right| + \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

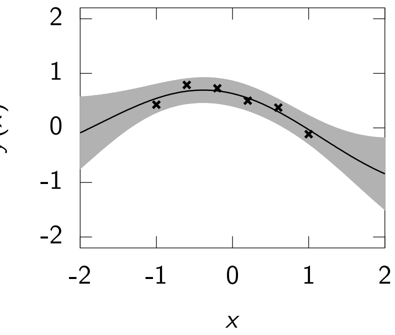
Can we determine length scales and noise levels from the data?

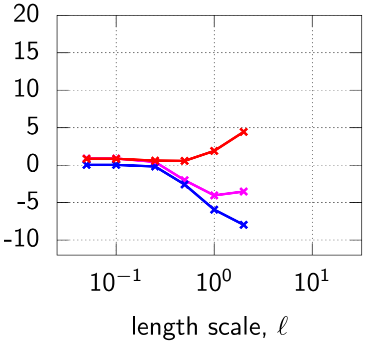


$$E(\boldsymbol{\theta}) = \frac{1}{2}\left|\mathbf{K}\right| + \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

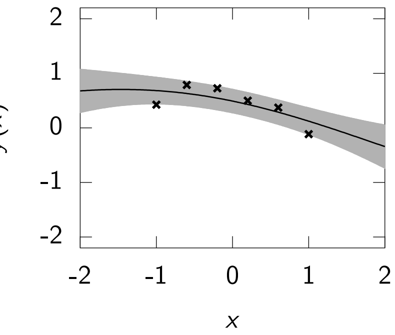


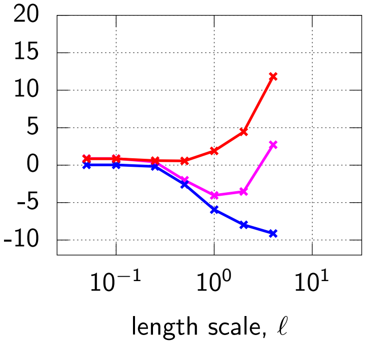


$$E(\boldsymbol{\theta}) = \frac{1}{2}\left|\mathbf{K}\right| + \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

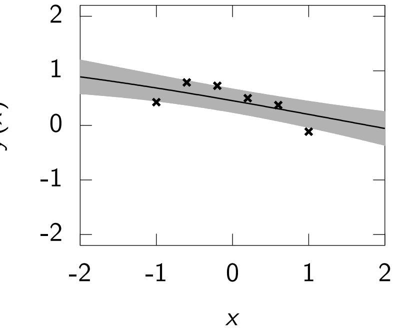


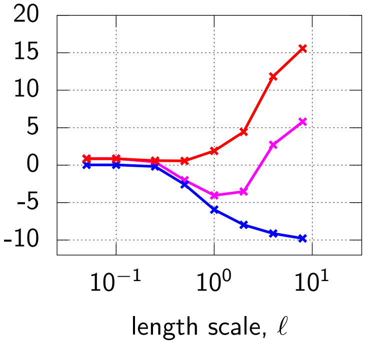


$$E(\boldsymbol{\theta}) = \frac{1}{2}\left|\mathbf{K}\right| + \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

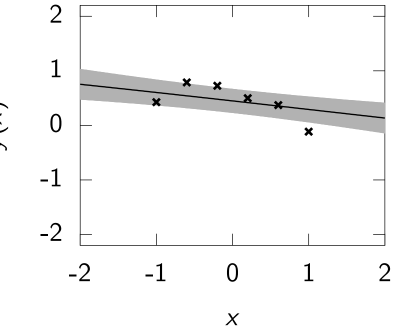


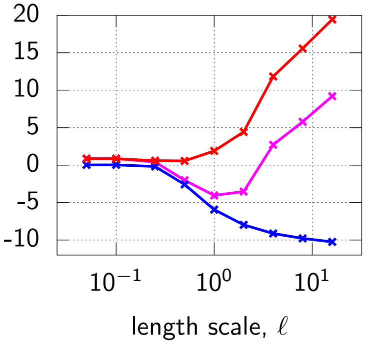


$$E(\boldsymbol{\theta}) = \frac{1}{2}\left|\mathbf{K}\right| + \frac{\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Covariance Parameters

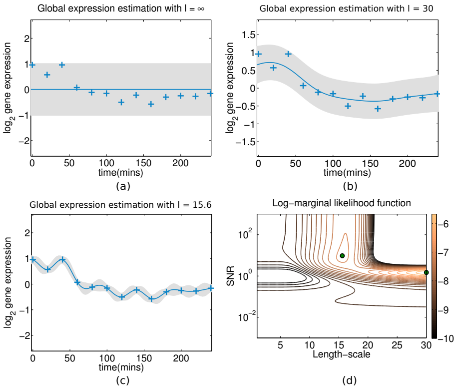
Can we determine length scales and noise levels from the data?





$$E(\boldsymbol{\theta}) = \frac{1}{2} |\mathbf{K}| + \frac{\mathbf{y}^\top \mathbf{K}^{-1} \mathbf{y}}{2}$$

Gene Expression Example



Data from **?**. Figure from **?**.

Outline

Limitations of Gaussian Processes

- Inference is $O(n^3)$ due to matrix inverse (in practice use Cholesky).
- Gaussian processes don't deal well with discontinuities (financial crises, phosphorylation, collisions, edges in images).
- Widely used exponentiated quadratic covariance (RBF) can be too smooth in practice (but there are many alternatives!!).

Summary

- Broad introduction to Gaussian processes.
  - Started with Gaussian distribution.
  - Motivated Gaussian processes through the multivariate density.
- Emphasized the role of the covariance (not the mean).
- Performs nonlinear regression with error bars.
- Parameters of the covariance function (kernel) are easily optimized with maximum likelihood.

References I

G. Della Gatta, M. Bansal, A. Ambesi-Impiombato, D. Antonini, C. Missero, and D. di Bernardo. Direct targets of the trp63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. *Genome Research*, 18(6): 939–948, Jun 2008. [URL]. [DOI].

A. A. Kalaitzis and N. D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, 12(180), 2011. [DOI].

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996. Lecture Notes in Statistics 118.

J. Oakley and A. O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4): 769–784, 2002.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. [Google Books] .

C. K. I. Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998.